The Indispensibility of Neutrality in Molecular Evolution

Peter Schuster

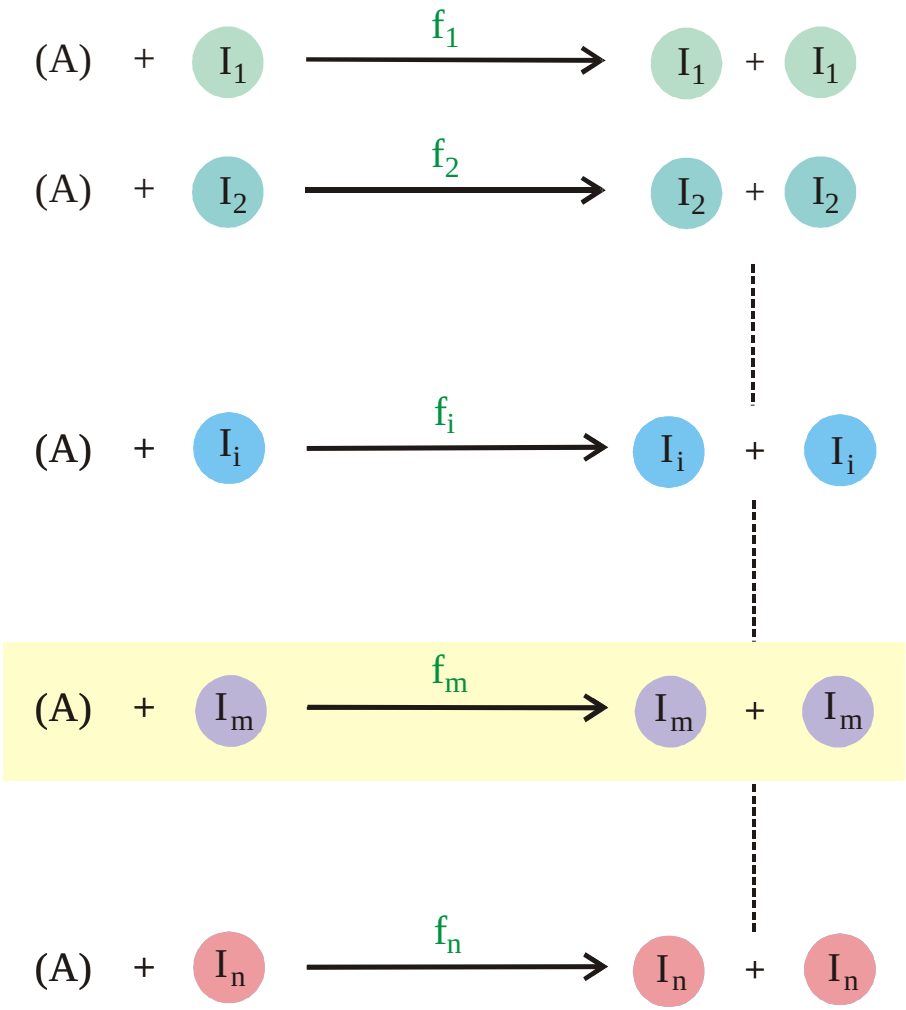
Institut für Theoretische Chemie und Molekulare Strukturbiologie der Universität Wien

Max-Planck-Institut für Physik Komplexer Systeme

Dresden, 05.07.2004

Web-Page for further information:

http://www.tbi.univie.ac.at/~pks



$$dx_i / dt = f_i x_i - x_i \Phi = x_i (f_i - \Phi)$$

$$\Phi = \Sigma_j f_j x_j ; \quad \Sigma_j x_j = 1 ; \quad i,j = 1,2,...,n$$

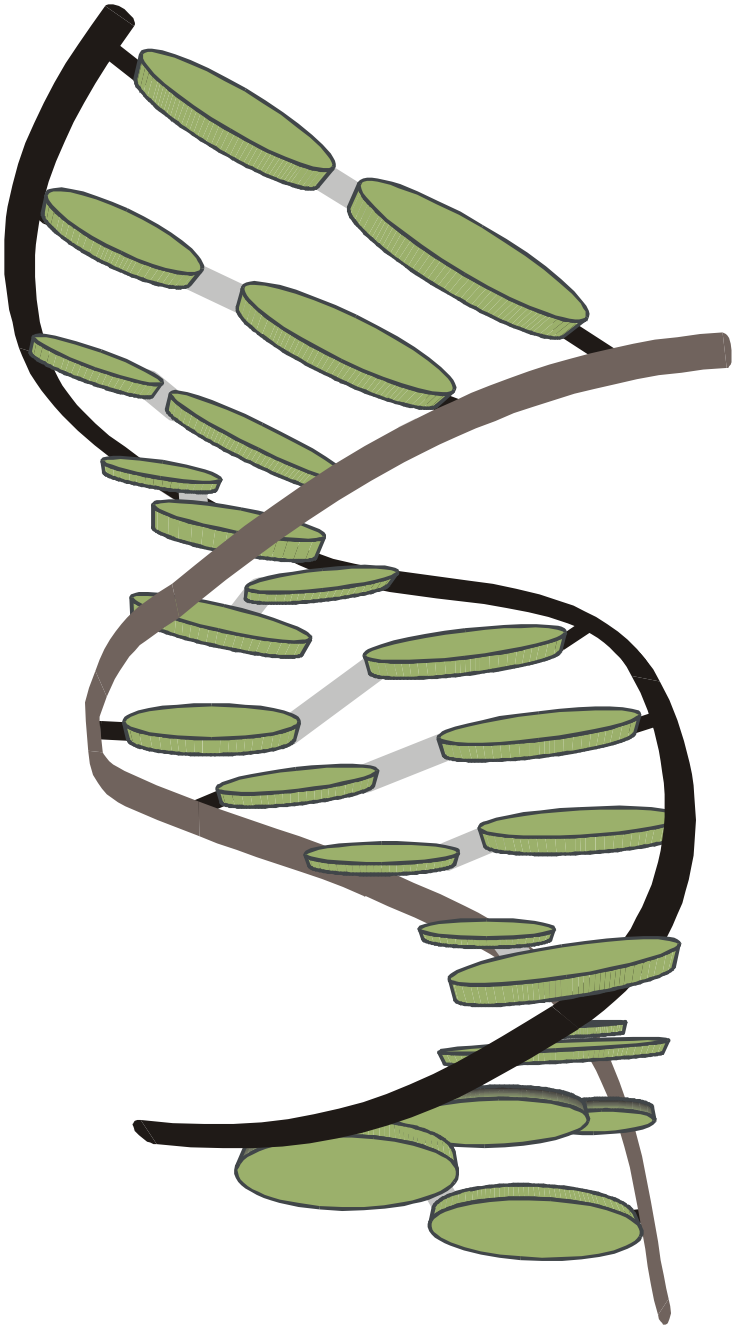
$$[I_i] = x_i \geq 0 ; \quad i = 1,2,...,n ;$$

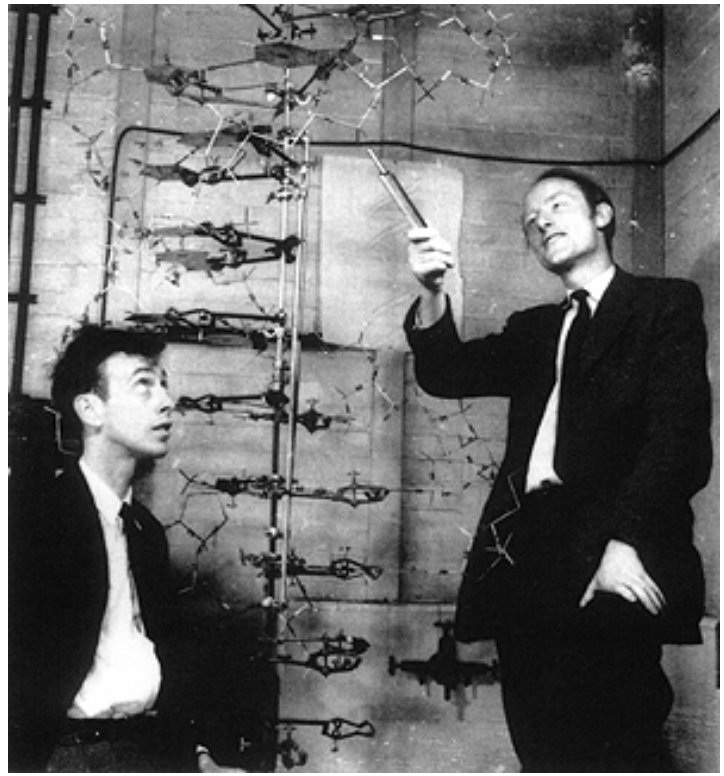
$$[A] = a = \text{constant}$$

$$f_m = \max \{f_j; j=1,2,...,n\}$$

$$x_m(t) \rightarrow 1 \quad \text{for} \quad t \rightarrow \infty$$

Reproduction of organisms **or replication** of molecules as the basis of selection





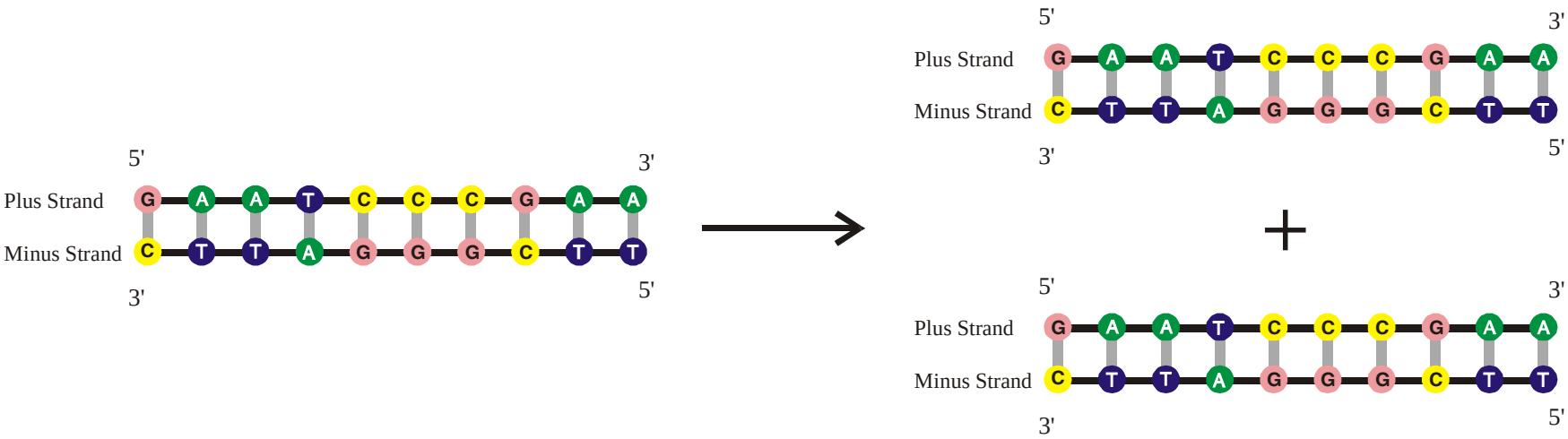
James D. Watson and Francis H.C. Crick

Nobel prize 1962

1953 – 2003 fifty years double helix

Base pairs: **A** = **T** and **G** ≡ **C**

Stacking of base pairs in nucleic acid double helices (B-DNA)



Direct replication of DNA is a higly complex copying mechanism involving more than ten different protein molecules. Complementarity is determined by Watson-Crick base pairs:

A=T and G≡C

Selection equation: $[I_i] = x_i \geq 0\ ,\ f_i > 0$

$$\frac{dx_i}{dt} = x_i\left(f_i - \phi\right), \quad i=1,2,\cdots,n; \quad \sum_{i=1}^{n} x_i = 1; \quad \phi = \sum_{j=1}^{n} f_j x_j = \overline{f}$$

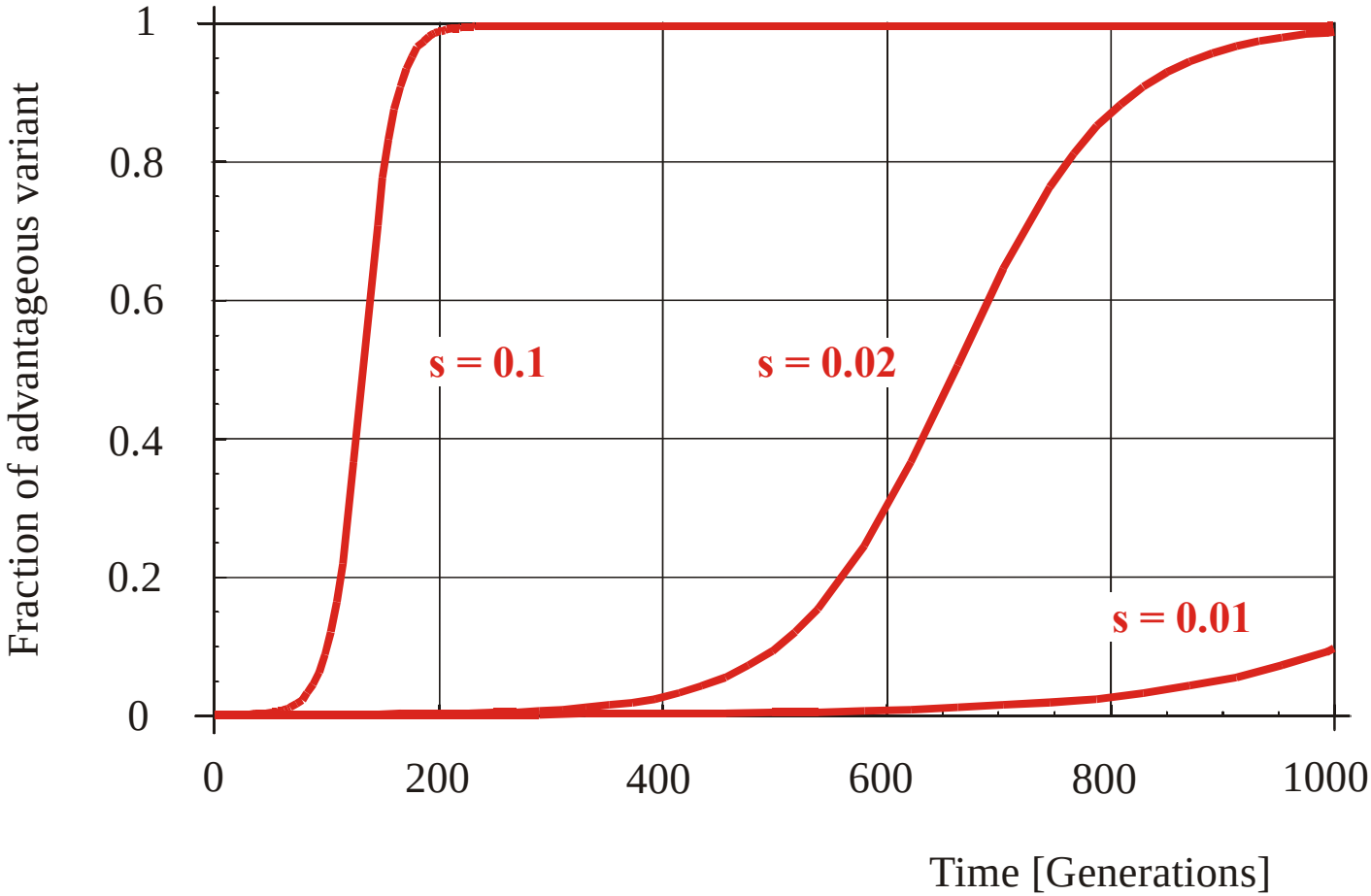
Mean fitness or dilution flux, ϕ(t), is a **non-decreasing function** of time,

$$\frac{d\phi}{dt} = \sum_{i=1}^{n} f_i \frac{dx_i}{dt} = \overline{f^2} - \left(\overline{f}\right)^2 = \mathrm{var}\{f\} \geq 0$$

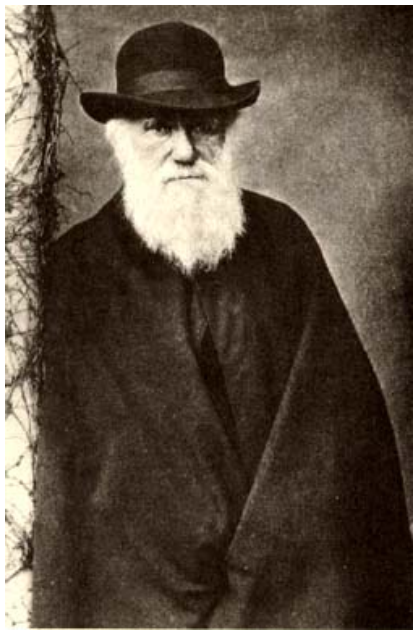
Solutions are obtained by integrating factor transformation

$$x_i(t) = \frac{x_i(0)\cdot\exp(f_i t)}{\sum_{j=1}^{n} x_j(0)\cdot\exp(f_j t)}; \quad i=1,2,\cdots,n$$

$s = (f_2 - f_1) / f_1;\ f_2 > f_1\ ;\ x_1(0) = 1 - 1/N\ ;\ x_2(0) = 1/N$



Selection of advantageous mutants in populations of N = 10 000 individuals



„…Variations neither useful not injurious would not be affected by natural selection, and would be left either a fluctuating element, as perhaps we see in certain polymorphic species, or would ultimately become fixed, owing to the nature of the organism and the nature of the conditions. …“

Charles Darwin, Origin of species (1859)

MOTOO KIMURA

Motoo Kimura's population genetics of neutral evolution.

Evolutionary rate at the molecular level. *Nature* **217**: 624-626, 1955.

The Neutral Theory of Molecular Evolution. Cambridge University Press. Cambridge, UK, 1983.

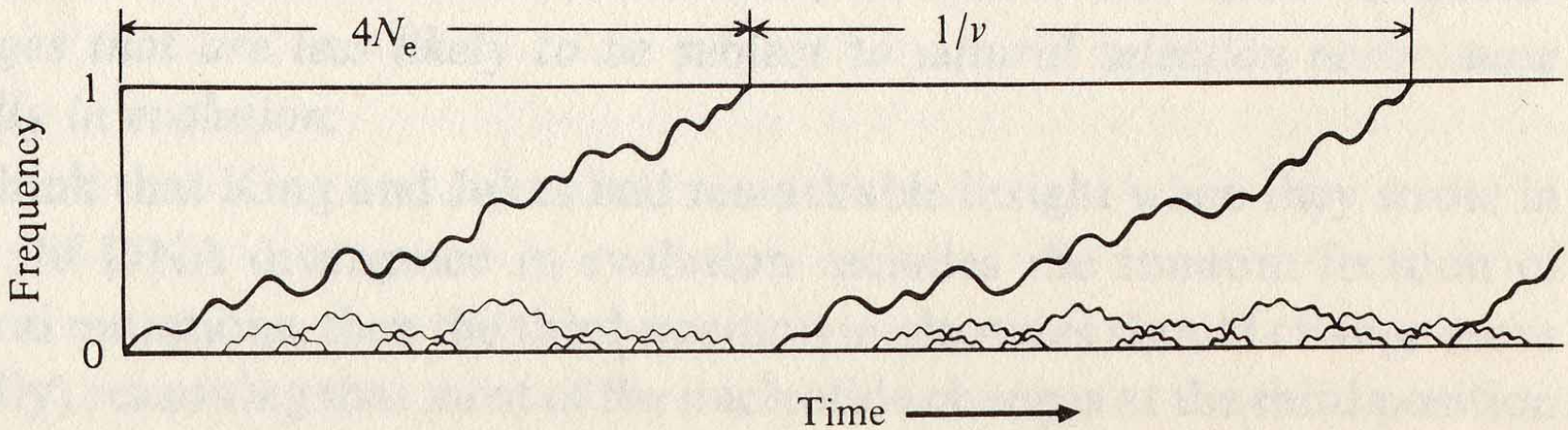
THE NEUTRAL THEORY
OF MOLECULAR EVOLUTION

MOTOO KIMURA
National Institute of Genetics, Japan

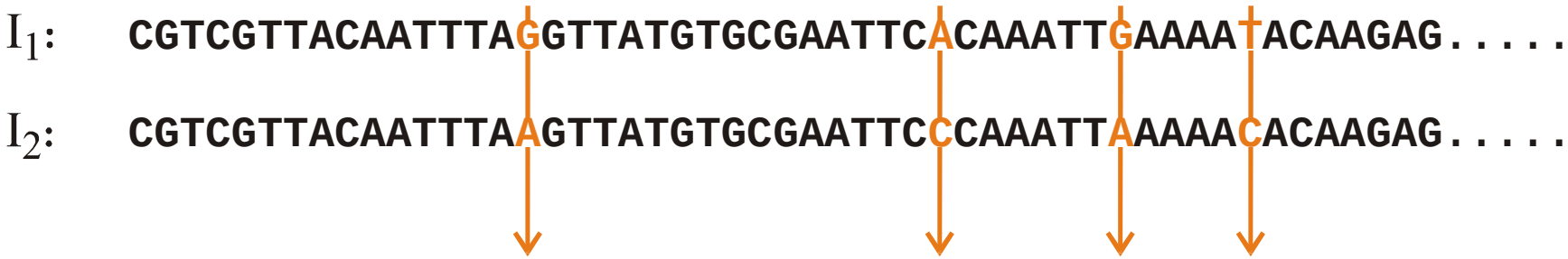
The right of the University of Cambridge to print and sell all manner of books was granted by Henry VIII in 1534. The University has printed and published continuously since 1584.

CAMBRIDGE UNIVERSITY PRESS
Cambridge
London New York New Rochelle
Melbourne Sydney

Fig. 3.1. Behavior of mutant genes following their appearance in a finite population. Courses of change in the frequencies of mutants destined to fixation are depicted by thick paths. N_e stands for the effective population size and v is the mutation rate.



Motoo Kimura. *The Neutral Theory of Molecular Evolution*.
Cambridge University Press. Cambridge, UK, 1983.



Hamming distance $d_H(I_1,I_2) = 4$

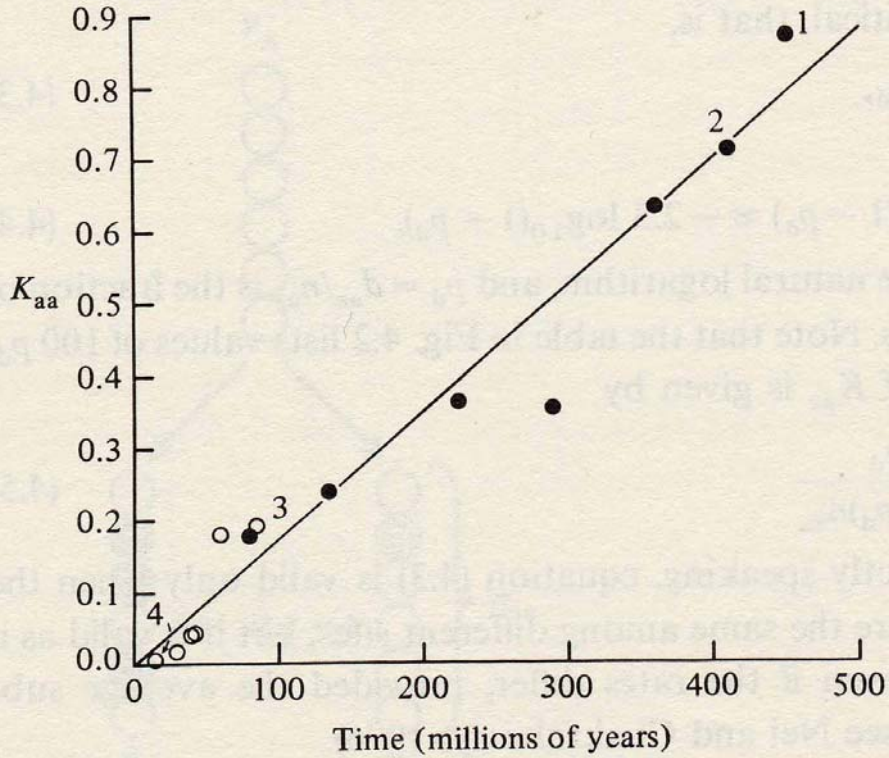
(i) $d_H(I_1,I_1) = 0$

(ii) $d_H(I_1,I_2) = d_H(I_2,I_1)$

(iii) $d_H(I_1,I_3) \leq d_H(I_1,I_2) + d_H(I_2,I_3)$

The Hamming distance between genotypes induces a metric in sequence space

Fig. 4.4. Relationship between K_{aa}, the number of amino acid substitutions (ordinate) and the time of divergence in millions of years (abscissa).

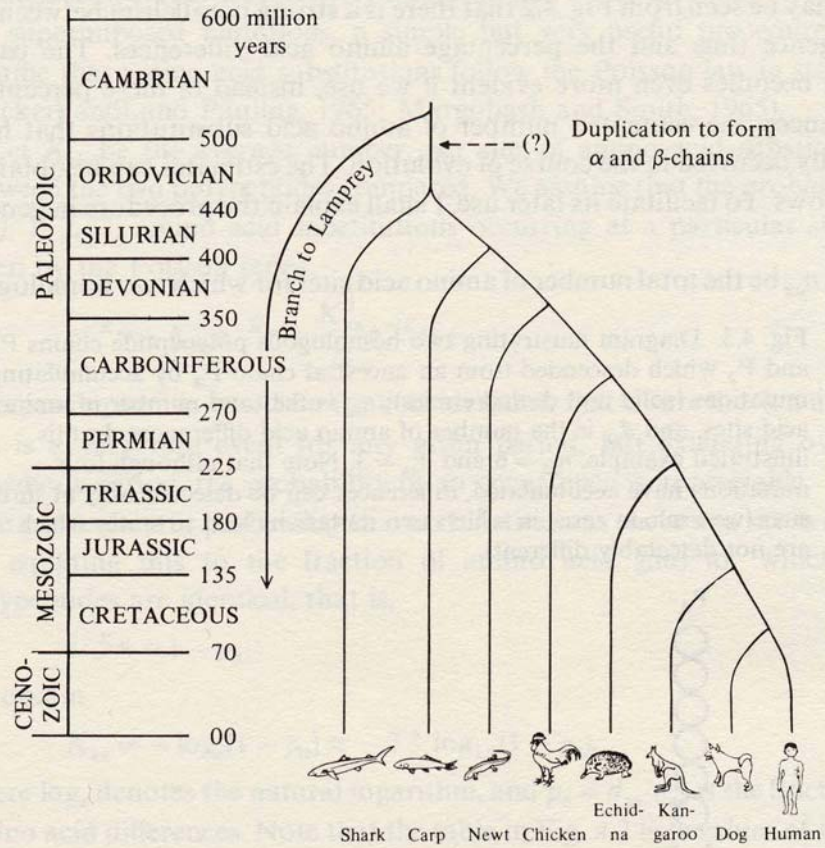


Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press. Cambridge, UK, 1983.

The molecular clock of evolution

Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press. Cambridge, UK, 1983.

Fig. 4.2. Percentage amino acid differences when the α hemoglobin chains are compared among eight vertebrates together with their phylogenetic relationship and the times of divergence.



	Shark	Carp	Newt	Chicken	Echidna	Kangaroo	Dog	Human
Shark		59.4	61.4	59.7	60.4	55.4	56.8	53.2
Carp			53.2	51.4	53.6	50.7	47.9	48.6
Newt				44.7	50.4	47.5	46.1	44.0
Chicken					34.0	29.1	31.2	24.8
Echidna						34.8	29.8	26.2
Kangaroo							23.4	19.1
Dog								16.3
Human								

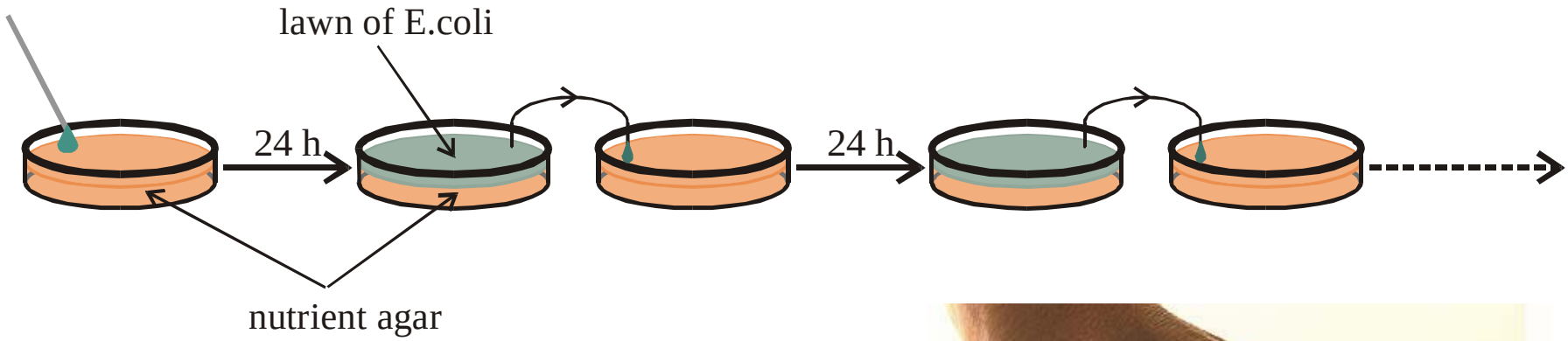
α-chain

Percent Amino Acid Differences

Bacterial Evolution

S. F. Elena, V. S. Cooper, R. E. Lenski. ***Punctuated evolution caused by selection of rare beneficial mutants***. Science **272** (1996), 1802-1804

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. ***Genomic evolution during a 10,000-generation experiment with bacteria***. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812



Serial transfer of Escherichia coli cultures in Petri dishes

1 day ≈ 6.67 generations
1 month ≈ 200 generations
1 year ≈ 2400 generations

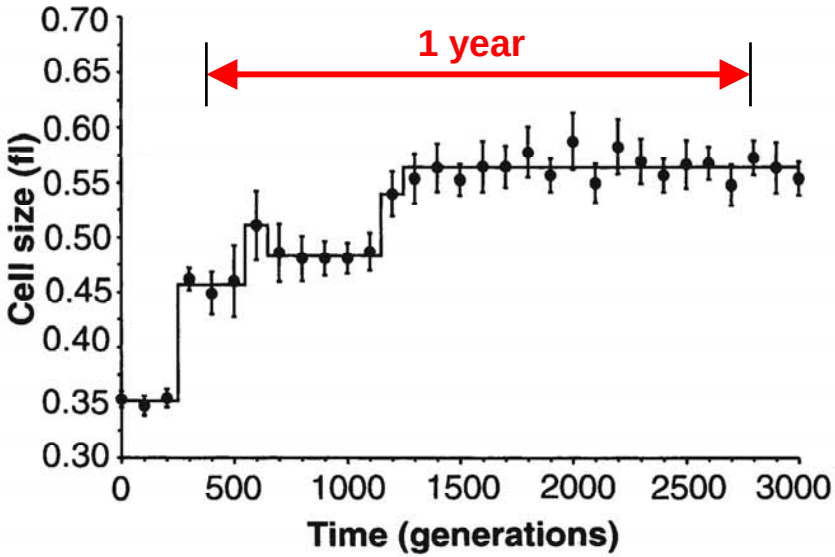


Fig. 1. Change in average cell size (1 fl = 10^{-15} L) in a population of *E. coli* during 3000 generations of experimental evolution. Each point is the mean of 10 replicate assays (*22*). Error bars indicate 95% confidence intervals. The solid line shows the best fit of a step-function model to these data (Table 1).

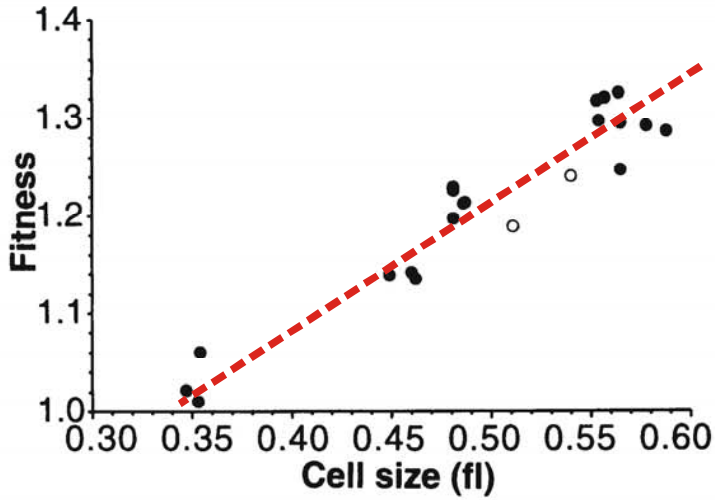
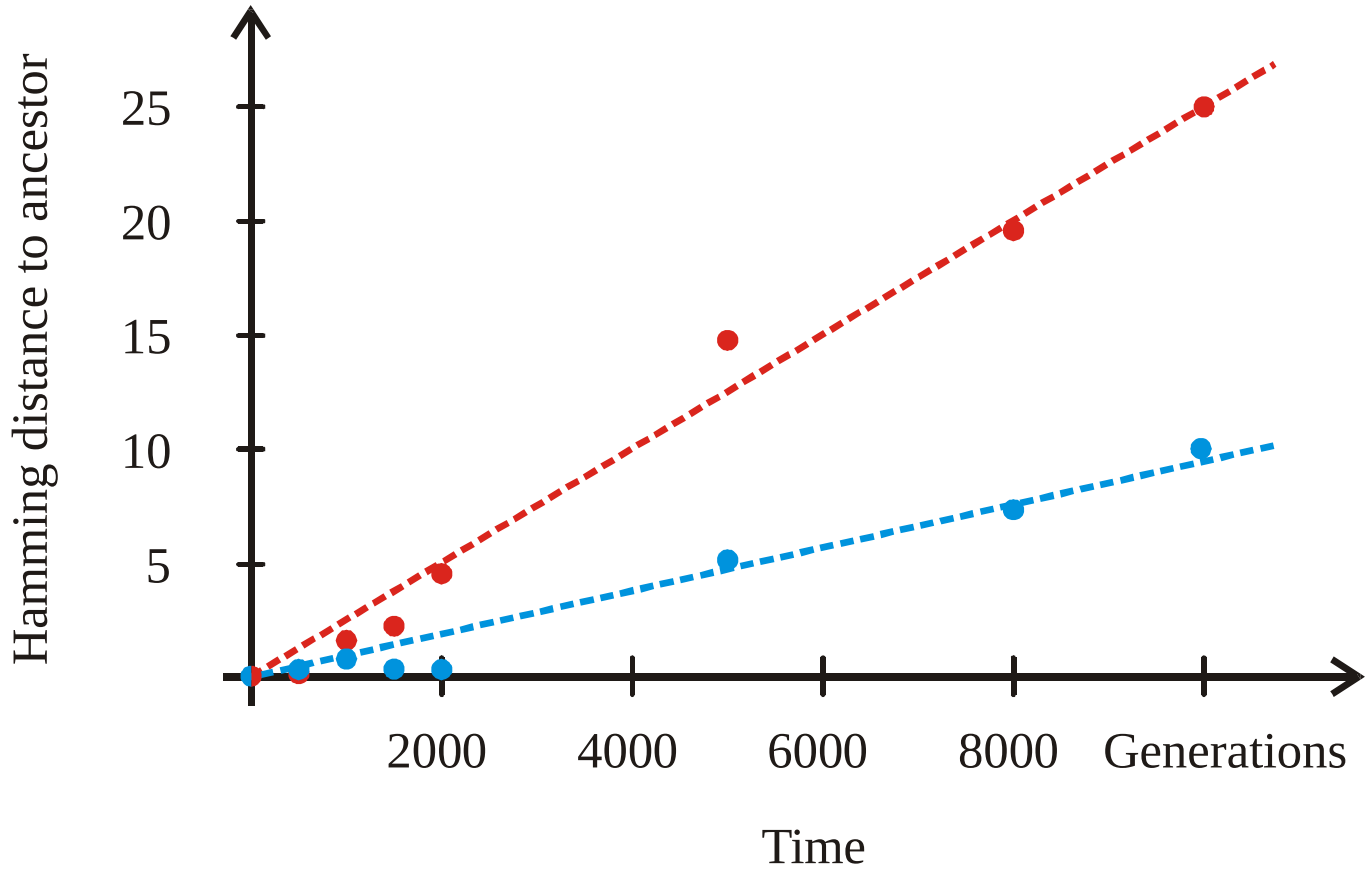


Fig. 2. Correlation between average cell size and mean fitness, each measured at 100-generation intervals for 2000 generations. Fitness is expressed relative to the ancestral genotype and was obtained from competition experiments between derived and ancestral cells (*6, 7*). The open symbols indicate the only two samples assigned to different steps by the cell size and fitness data.

Epochal evolution of bacteria in serial transfer experiments under constant conditions

S. F. Elena, V. S. Cooper, R. E. Lenski. *Punctuated evolution caused by selection of rare beneficial mutants*. Science **272** (1996), 1802-1804

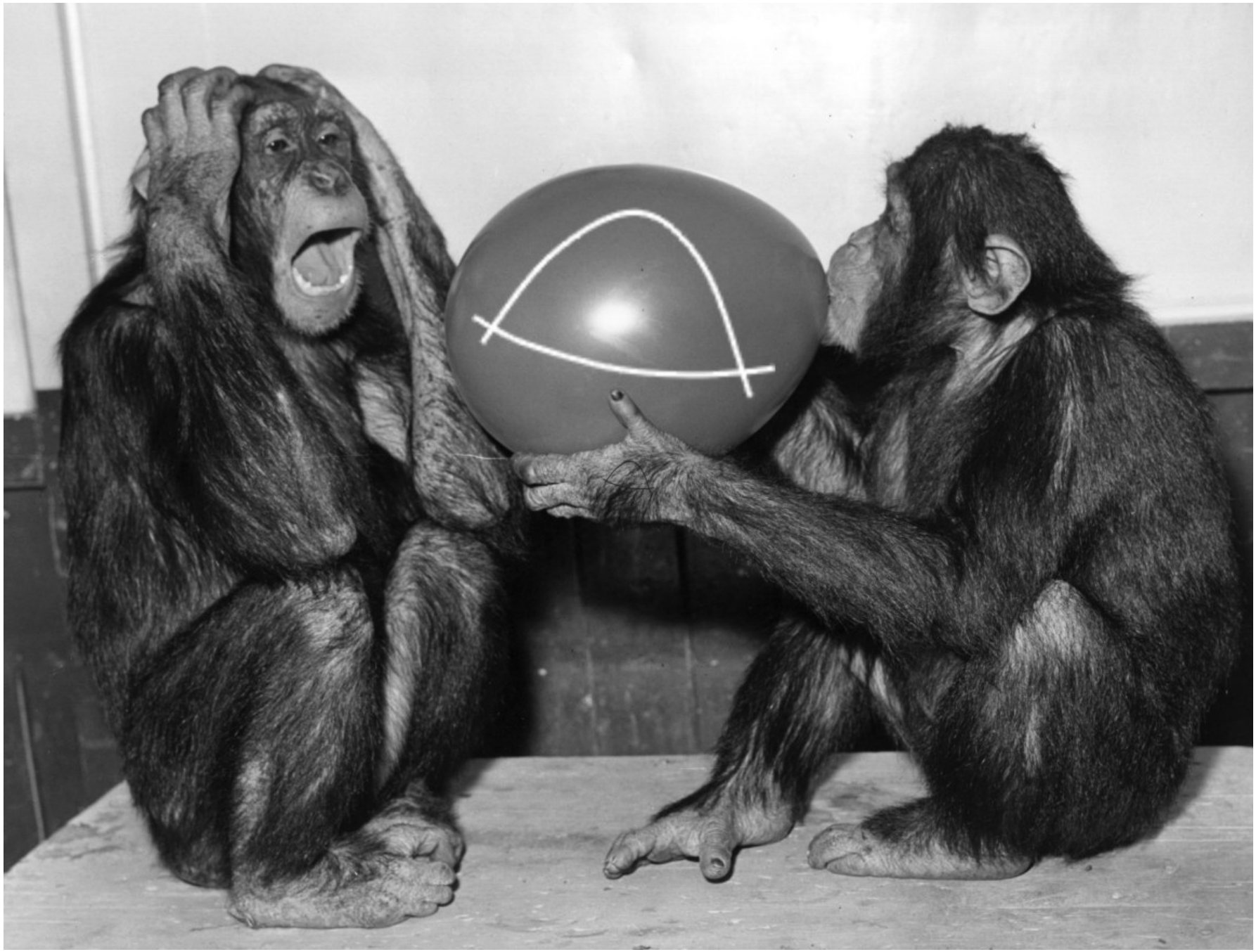


Variation of genotypes in a bacterial serial transfer experiment

D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, M. Blot. *Genomic evolution during a 10,000-generation experiment with bacteria*. Proc.Natl.Acad.Sci.USA **96** (1999), 3807-3812

No new principle will declare itself from below a heap of facts.

Sir Peter Medawar, 1985



In evolution **variation** occurs on **genotypes** but **selection** operates on the **phenotype**.

Mappings from genotypes into phenotypes are highly complex objects. The only computationally accessible case is in the evolution of RNA molecules.

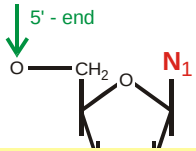
The mapping from RNA sequences into secondary structures and function,

$$\textbf{sequence} \Rightarrow \textbf{structure} \Rightarrow \textbf{fitness},$$

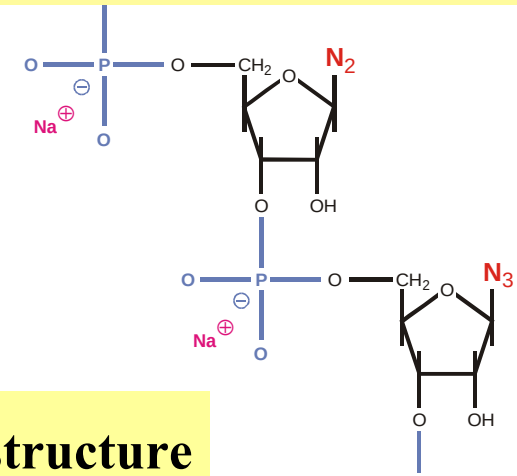
is used as a model for the complex relations between genotypes and phenotypes. Fertile progeny measured in terms of **fitness** in population biology is determined quantitatively by **replication rate constants** of RNA molecules.

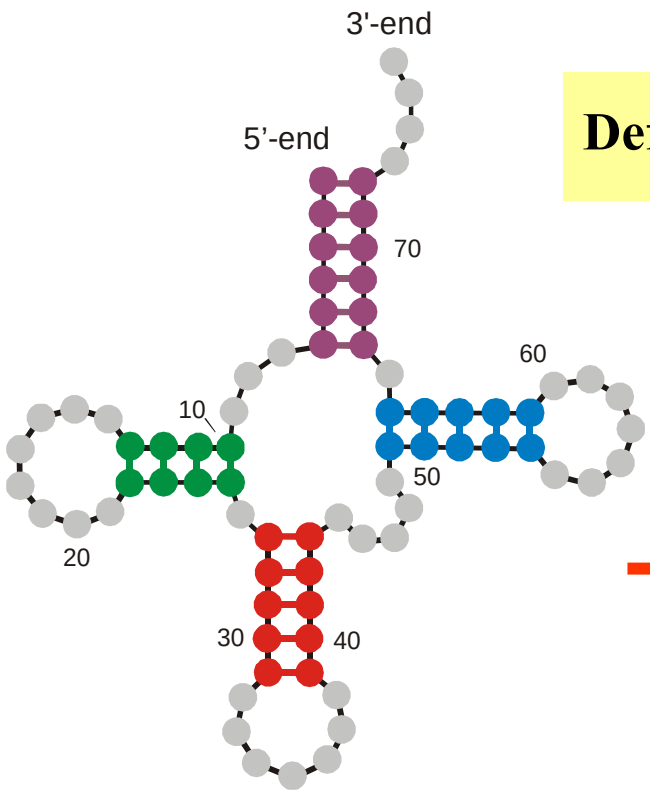
Population biology	Molecular genetics	Evolution of RNA molecules
Genotype	**Genome**	**RNA sequence**
Phenotype	**Organism**	**RNA structure and function**
Fitness	**Reproductive success**	**Replication rate constant**

The RNA model

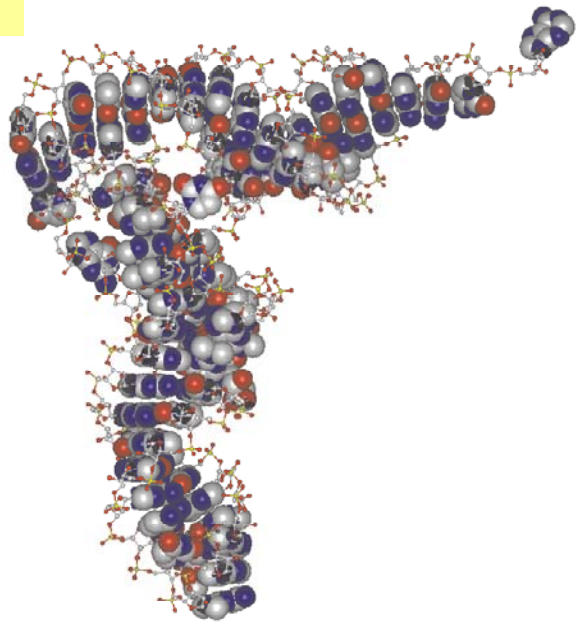


5'-end GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA 3'-end





Definition of RNA structure

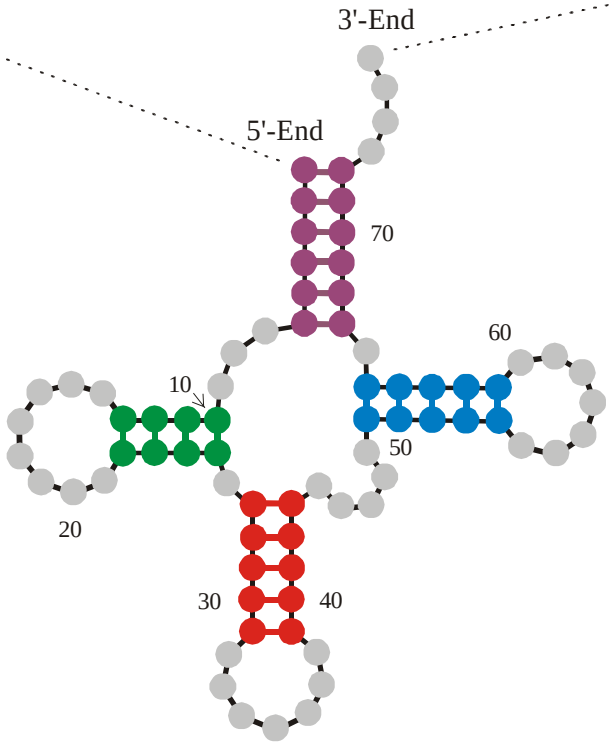


Sequence

5'-End 3'-End

GCGGAUUUAGCUCAGDDGGGAGAGCMCCAGACUGAAYAUCUGGAGMUCCUGUGTPCGAUCCACAGAAUUCGCACCA

Secondary structure



Base pairs: A = U , G ≡ C , and G - U

Definition and **physical relevance** of RNA secondary structures

RNA secondary structures are listings of Watson-Crick and GU wobble base pairs, which are free of knots and pseudokots.

D.Thirumalai, N.Lee, S.A.Woodson, and D.K.Klimov.
Annu.Rev.Phys.Chem. **52**:751-762 (2001):

„**Secondary structures are folding intermediates in the formation of full three-dimensional structures**.“

RNA sequence GUAUCGAAAUACGUAGCGUAUGGGGAUGCUGGACGGUCCCAUCGGUACUCCA

RNA folding:

Structural biology, spectroscopy of biomolecules, understanding **molecular function**

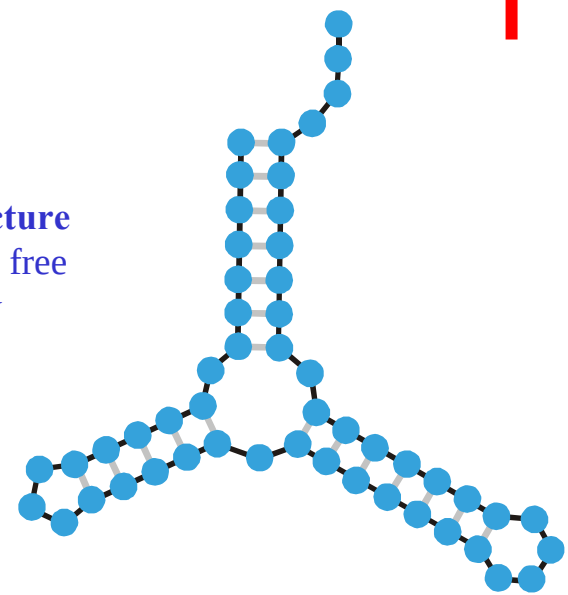
Biophysical chemistry: thermodynamics and kinetics

Empirical parameters

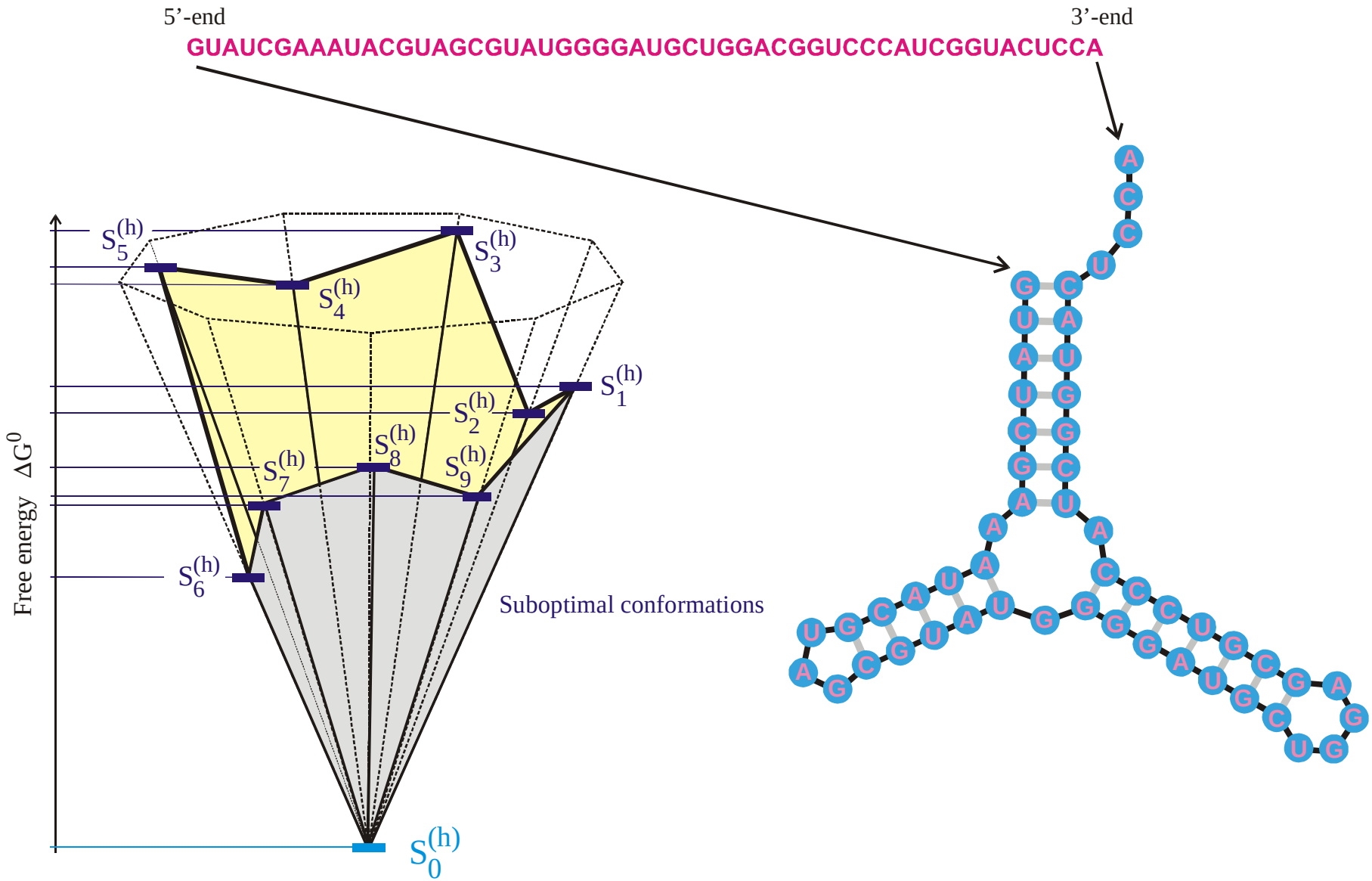
Inverse folding of RNA:

Biotechnology, **design of biomolecules** with predefined structures and functions

RNA structure of minimal free energy



Sequence, structure, and design



The minimum free energy structures on a discrete space of conformations

Criterion of Minimum Free Energy

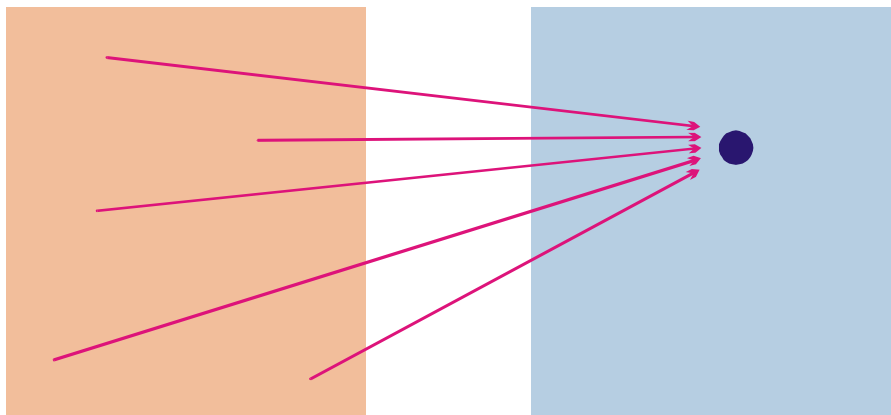
UUUAGCCAGCGCGAGUCGUGCGGACGGGGUUAUCUCUGUCGGGCUAGGGCGC

GUGAGCGCGGGGCACAGUUUCUCAAGGAUGUAAGUUUUUGCCGUUUAUCUGG

UUAGCGAGAGAGGAGGCUUCUAGACCCAGCUCUCUGGGUCGUUGCUGAUGCG

CAUUGGUGCUAAUGAUAUUAGGGCUGUAUUCCUGUAUAGCGAUCAGUGUCCG

GUAGGCCCUCUUGACAUAAGAUUUUUCCAAUGGUGGGAGAUGGCCAUUGCAG



Sequence Space

Shape Space

From sequences to shapes and back: a case study in RNA secondary structures

PETER SCHUSTER[1,2,3], WALTER FONTANA[3], PETER F. STADLER[2,3] AND IVO L. HOFACKER[2]

[1] *Institut für Molekulare Biotechnologie, Beutenbergstrasse 11, PF 100813, D-07708 Jena, Germany*
[2] *Institut für Theoretische Chemie, Universität Wien, Austria*
[3] *Santa Fe Institute, Santa Fe, U.S.A.*

SUMMARY

RNA folding is viewed here as a map assigning secondary structures to sequences. At fixed chain length the number of sequences far exceeds the number of structures. Frequencies of structures are highly non-uniform and follow a generalized form of Zipf's law: we find relatively few common and many rare ones. By using an algorithm for inverse folding, we show that sequences sharing the same structure are distributed randomly over sequence space. All common structures can be accessed from an arbitrary sequence by a number of mutations much smaller than the chain length. The sequence space is percolated by extensive neutral networks connecting nearest neighbours folding into identical structures. Implications for evolutionary adaptation and for applied molecular evolution are evident: finding a particular structure by mutation and selection is much simpler than expected and, even if catalytic activity should turn out to be sparse in the space of RNA structures, it can hardly be missed by evolutionary processes.

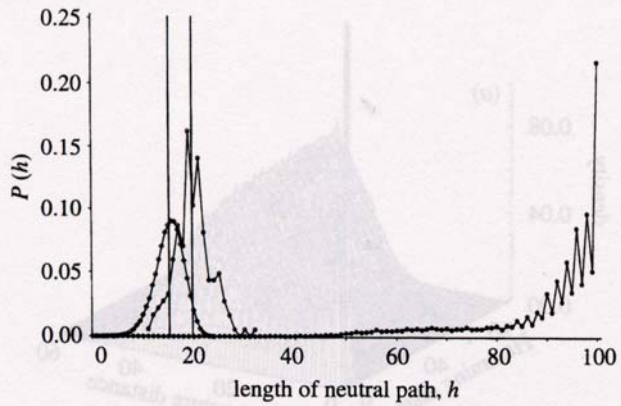


Figure 4. Neutral paths. A neutral path is defined by a series of nearest neighbour sequences that fold into identical structures. Two classes of nearest neighbours are admitted: neighbours of Hamming distance 1, which are obtained by single base exchanges in unpaired stretches of the structure, and neighbours of Hamming distance 2, resulting from base pair exchanges in stacks. Two probability densities of Hamming distances are shown that were obtained by searching for neutral paths in sequence space: (i) an upper bound for the closest approach of trial and target sequences (open circles) obtained as endpoints of neutral paths approaching the target from a random trial sequence (185 targets and 100 trials for each were used); (ii) a lower bound for the closest approach of trial and target sequences (open diamonds) derived from secondary structure statistics (Fontana *et al.* 1993*a*; see this paper, §4); and (iii) longest distances between the reference and the endpoints of monotonously diverging neutral paths (filled circles) (500 reference sequences were used).

Proc. R. Soc. Lond. B (1994) **255**, 279–284
Printed in Great Britain

Reference for postulation and ***in silico*** verification of ***neutral networks***

Evolution *in silico*

W. Fontana, P. Schuster,
Science **280** (1998), 1451-1455

random individuals. The primer pair used for genomic DNA amplification is 5′-TCTCCCTGGATTCT-CATTTA-3′ (forward) and 5′-TCTTTGTCTTCTGT-TCCACC-3′ (reverse). Reactions were performed in 25 μl using 1 unit of Taq DNA polymerase with each primer at 0.4 μM; 200 μM each dATP, dTTP, dGTP, and dCTP; and PCR buffer [10 mM tris-HCl (pH 8.3), 50 mM KCl_2,1.5 mM $MgCl_2$] in a cycle condition of 94°C for 1 min and then 35 cycles of 94°C for 30 s, 55°C for 30 s, and 72°C for 30 s followed by 72°C for 6 min. PCR products were purified (Qiagen), digested with Xmn I, and separated in a 2% agarose gel.

32. A nonsense mutation may affect mRNA stability and result in degradation of the transcript [L. Maquat, *Am. J. Hum. Genet.* **59**, 279 (1996)].
33. Data not shown; a dot blot with poly (A)$^+$ RNA from 50 human tissues (The Human RNA Master Blot, 7770-1, Clontech Laboratories) was hybridized with a probe from exons 29 to 47 of *MYO15* using the same condition as Northern blot analysis (*13*).
34. Smith–Magenis syndrome (SMS) is due to deletions of 17p11.2 of various sizes, the smallest of which includes *MYO15* and perhaps 20 other genes [(*6*); K-S Chen, L. Potocki, J. R. Lupski, *MRDD Res. Rev.* **2**, 122 (1996)]. *MYO15* expression is easily detected in the pituitary gland (data not shown). Haploinsufficiency for *MYO15* may explain a portion of the SMS phenotype such as short stature. Moreover, a few SMS patients have sensorineural hearing loss, possibly because of a point mutation in *MYO15* in trans to the SMS 17p11.2 deletion.
35. R. A. Fridell, data not shown.
36. K. B. Avraham *et al.*, *Nature Genet.* **11**, 369 (1995); X-Z. Liu *et al.*, *ibid.* **17**, 268 (1997); F. Gibson *et al.*, *Nature* **374**, 62 (1995); D. Weil *et al.*, *ibid.*, p. 60.
37. RNA was extracted from cochlea (membranous labyrinths) obtained from human fetuses at 18 to 22 weeks of development in accordance with guidelines established by the Human Research Committee at the Brigham and Women's Hospital. Only samples without evidence of degradation were pooled for poly (A)$^+$ selection over oligo(dT) columns. First-strand cDNA was prepared using an Advantage RT-for-PCR kit (Clontech Laboratories). A portion of the first-strand cDNA (4%) was amplified by PCR with Advantage cDNA polymerase mix (Clontech Laboratories) using human *MYO15*-specific oligonucleotide primers (forward, 5′-GCATGACCTGCCGGCTAAT-GGG-3′; reverse, 5′-CTCACGGCTTCTGCATGGT-GCTCGGCTGGC-3′). Cycling conditions were 40 s at 94°C; 40 s at 66°C (3 cycles), 60°C (5 cycles), and 55°C (29 cycles); and 45 s at 68°C. PCR products were visualized by ethidium bromide staining after fractionation in a 1% agarose gel. A 688-bp PCR product is expected from amplification of the human *MYO15* cDNA. Amplification of human genomic DNA with this primer pair would result in a 2903-bp fragment.

38. We are grateful to the people of Bengkala, Bali, and the two families from India. We thank J. R. Lupski and K.-S. Chen for providing the human chromosome 17 cosmid library. For technical and computational assistance, we thank N. Dietrich, M. Fergusson, A. Gupta, E. Sorbello, R. Torkzadeh, C. Varner, M. Walker, G. Bouffard, and S. Beckstrom-Sternberg (National Institutes of Health Intramural Sequencing Center). We thank J. T. Hinnant, I. N. Arhya, and S. Winata for assistance in Bali, and T. Barber, S. Sullivan, E. Green, D. Drayna, and J. Battey for helpful comments on this manuscript. Supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) (Z01 DC 00035-01 and Z01 DC 00038-01 to T.B.F. and E.R.W. and R01 DC 03402 to C.C.M.), the National Institute of Child Health and Human Development (R01 HD30428 to S.A.C.) and a National Science Foundation Graduate Research Fellowship to F.J.P. This paper is dedicated to J. B. Snow Jr. on his retirement as the Director of the NIDCD.

9 March 1998; accepted 17 April 1998

Continuity in Evolution: On the Nature of Transitions

Walter Fontana and Peter Schuster

To distinguish continuous from discontinuous evolutionary change, a relation of nearness between phenotypes is needed. Such a relation is based on the probability of one phenotype being accessible from another through changes in the genotype. This nearness relation is exemplified by calculating the shape neighborhood of a transfer RNA secondary structure and provides a characterization of discontinuous shape transformations in RNA. The simulation of replicating and mutating RNA populations under selection shows that sudden adaptive progress coincides mostly, but not always, with discontinuous shape transformations. The nature of these transformations illuminates the key role of neutral genetic drift in their realization.

A much-debated issue in evolutionary biology concerns the extent to which the history of life has proceeded gradually or has been punctuated by discontinuous transitions at the level of phenotypes (*1*). Our goal is to make the notion of a discontinuous transition more precise and to understand how it arises in a model of evolutionary adaptation.

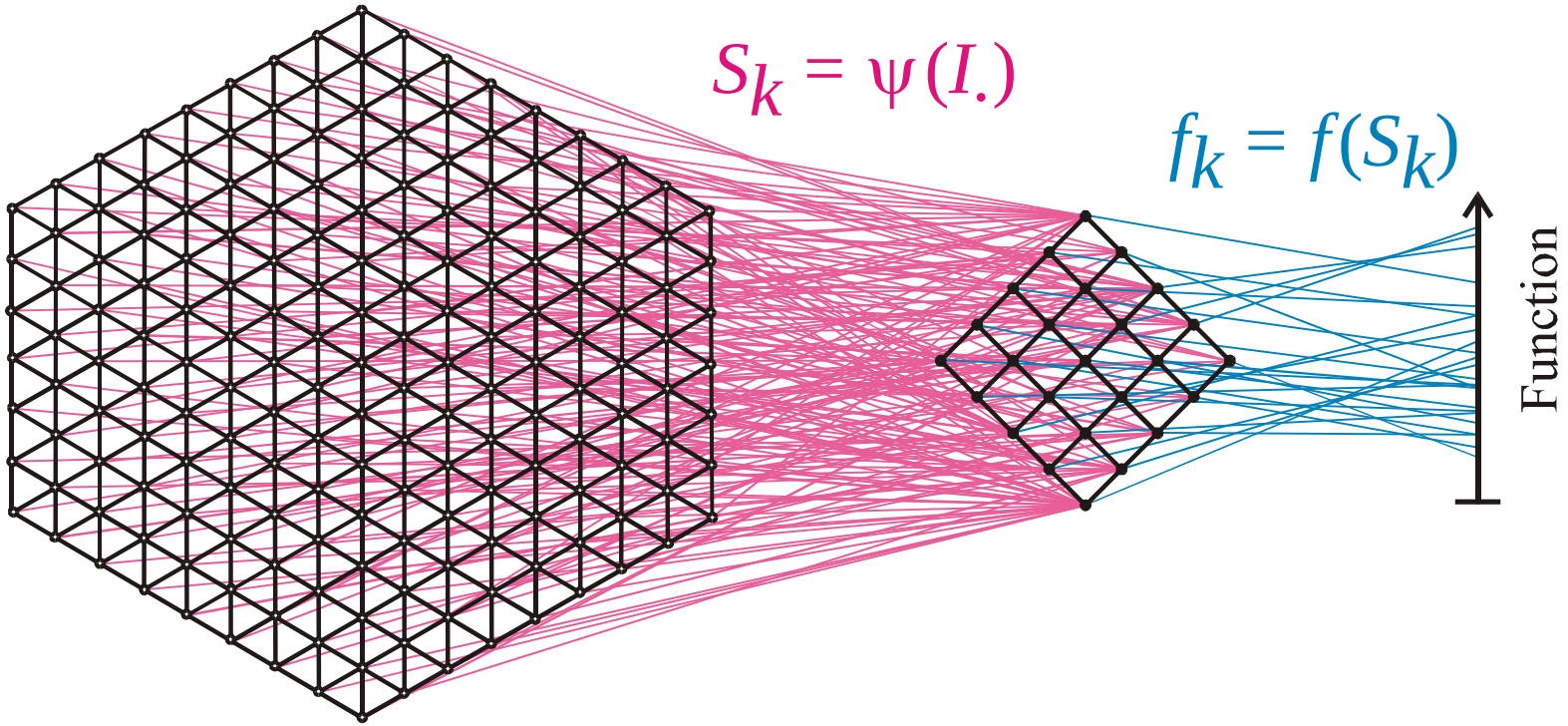
We focus on the narrow domain of RNA secondary structure, which is currently the simplest computationally tractable, yet realistic phenotype (*2*). This choice enables the definition and exploration of concepts that may prove useful in a wider context. RNA secondary structures represent a coarse level of analysis compared with the three-dimensional structure at atomic resolution. Yet, secondary structures are empirically well defined and obtain their biophysical and biochemical importance from being a scaffold for the tertiary structure. For the sake of brevity, we shall refer to secondary structures as "shapes." RNA combines in a single molecule both genotype (replicatable sequence) and phenotype (selectable shape), making it ideally suited for in vitro evolution experiments (*3*, *4*).

To generate evolutionary histories, we used a stochastic continuous time model of an RNA population replicating and mutating in a capacity-constrained flow reactor under selection (*5*, *6*). In the laboratory, a goal might be to find an RNA aptamer binding specifically to a molecule (*4*). Although in the experiment the evolutionary end product was unknown, we thought of its shape as being specified implicitly by the imposed selection criterion. Because our intent is to study evolutionary histories rather than end products, we defined a target shape in advance and assumed the replication rate of a sequence to be a function of the similarity between its shape and the target. An actual situation may involve more than one best shape, but this does not affect our conclusions.

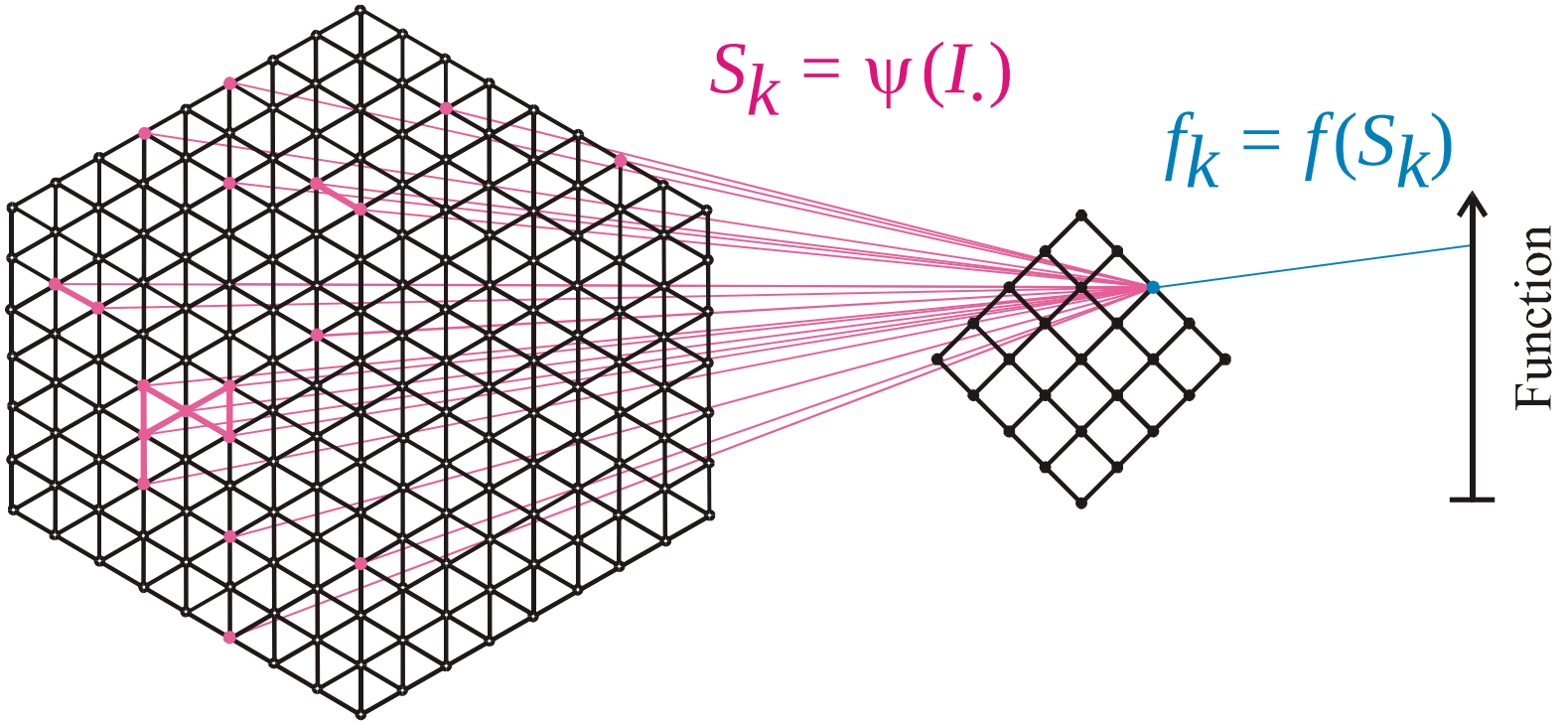
An instance representing in its qualitative features all the simulations we performed is shown in Fig. 1A. Starting with identical sequences folding into a random shape, the simulation was stopped when the population became dominated by the target, here a canonical tRNA shape. The black curve traces the average distance to the target (inversely related to fitness) in the population against time. Aside from a short initial phase, the entire history is dominated by steps, that is, flat periods of no apparent adaptive progress, interrupted by sudden approaches toward the target structure (*7*). However, the dominant shapes in the population not only change at these marked events but undergo several fitness-neutral transformations during the periods of no apparent progress. Although discontinuities in the fitness trace are evident, it is entirely unclear when and on the basis of what the series of successive phenotypes itself can be called continuous or discontinuous.

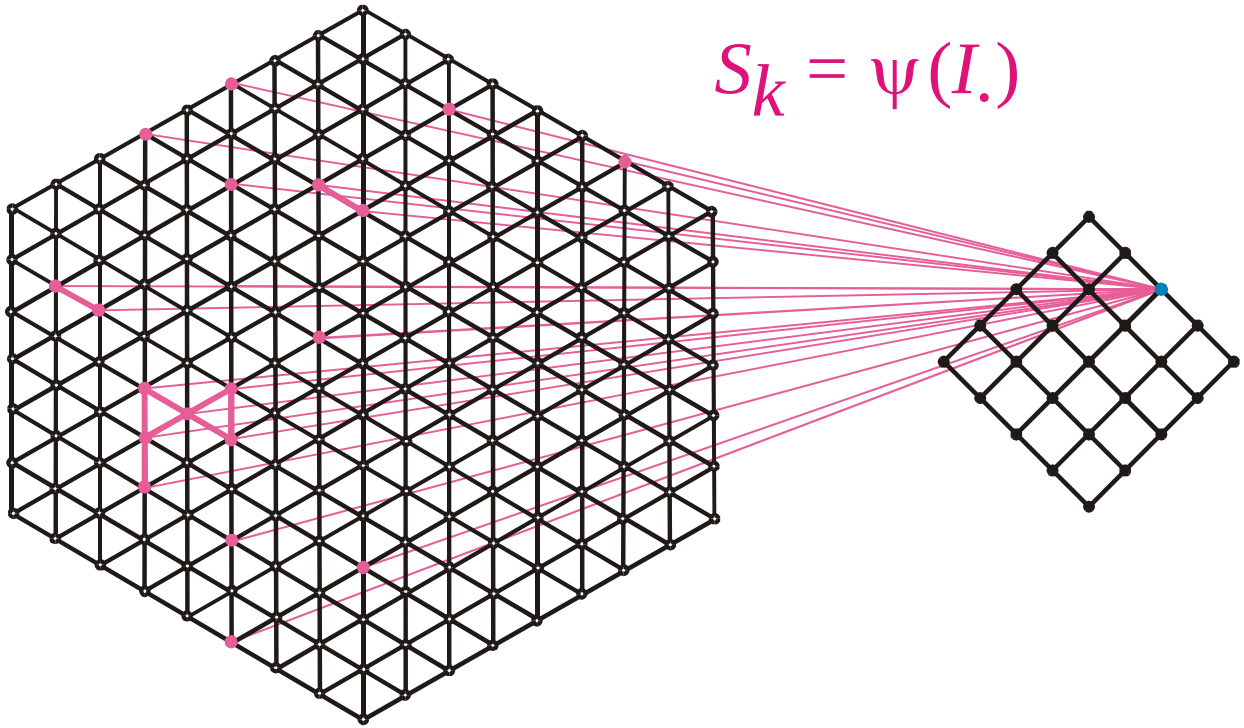
A set of entities is organized into a (topological) space by assigning to each entity a system of neighborhoods. In the present case, there are two kinds of entities: sequences and shapes, which are related by a thermodynamic folding procedure. The set of possible sequences (of fixed length) is naturally organized into a space because point mutations induce a canonical neighborhood. The neighborhood of a sequence consists of all its one-error mutants. The problem is how to organize the set of possible shapes into a space. The issue arises because, in contrast to sequences, there are

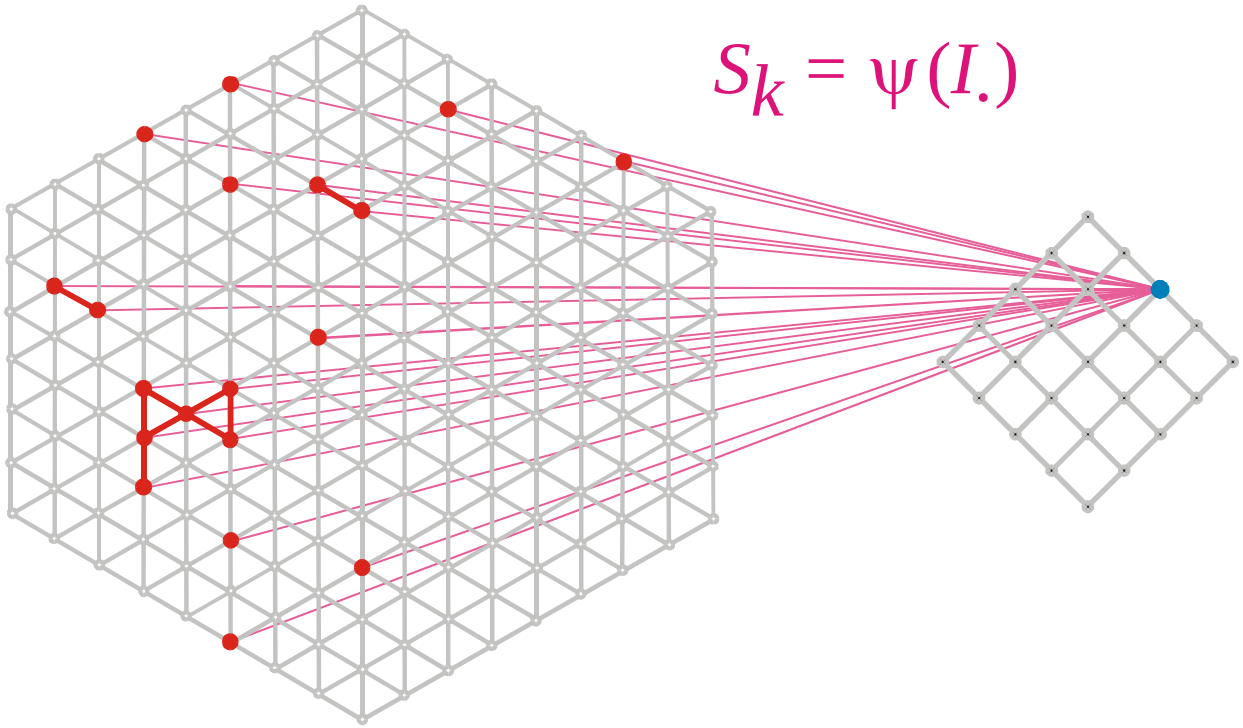
Institut für Theoretische Chemie, Universität Wien, Währingerstrasse 17, A-1090 Wien, Austria, Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA, and International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.



Mapping from sequence space into structure space and into function


$S_k = \psi(I_.)$
$f_k = f(S_k)$
Function
Sequence space
Structure space
Real numbers


$S_k = \psi(I_.)$
Sequence space
Structure space



The pre-image of the structure S_k in sequence space is the **neutral network G_k**

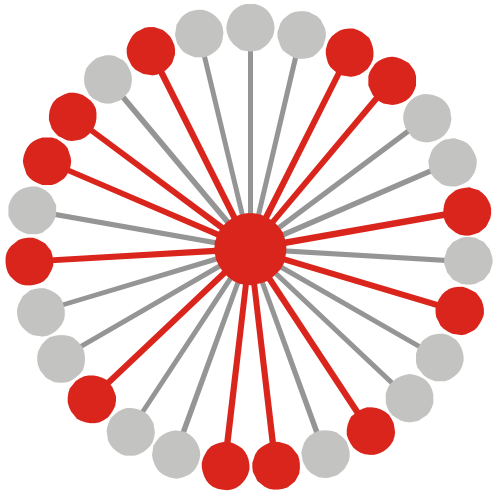
Neutral networks are sets of sequences forming the same object in a phenotype space. The neutral network $\mathbf{G_k}$ is, for example, the pre-image of the structure S_k in sequence space:

$$\mathbf{G_k} = \psi^{-1}(S_k) \pi \{\psi_j \mid \psi(I_j) = S_k\}$$

The set is converted into a graph by connecting all sequences of Hamming distance one.

Neutral networks of small biomolecules can be computed by exhaustive folding of complete sequence spaces, i.e. all RNA sequences of a given chain length. This number, $\mathbf{N=4^n}$, becomes very large with increasing length, and is prohibitive for numerical computations.

Neutral networks can be modelled by **random graphs** in sequence space. In this approach, nodes are inserted randomly into sequence space until the size of the pre-image, i.e. the number of neutral sequences, matches the neutral network to be studied.



$$G_k = \psi^{-1}(S_k) \cup \{ I_j \mid \psi(I_j) = S_k \}$$

$$\lambda_j = 12 / 27 = 0.444 , \quad \bar{\lambda}_k = \frac{\sum_{j \in |G_k|} \lambda_j(k)}{|G_k|}$$

Connectivity threshold: $\lambda_{cr} = 1 - \kappa^{-1/(\kappa-1)}$

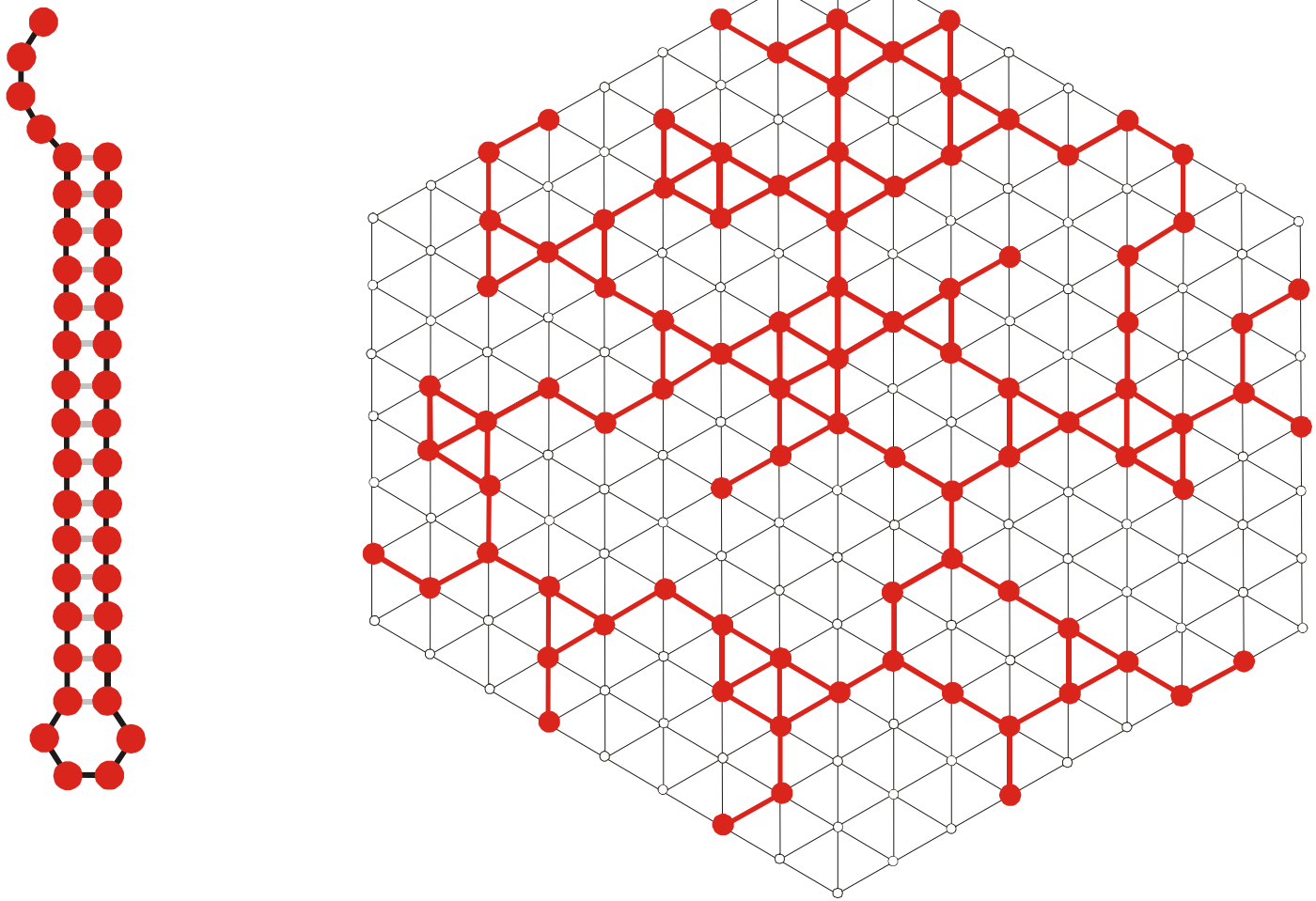
Alphabet size κ : **AUGC** $\Rightarrow \kappa = 4$

$\bar{\lambda}_k > \lambda_{cr}$ .... network G_k is connected

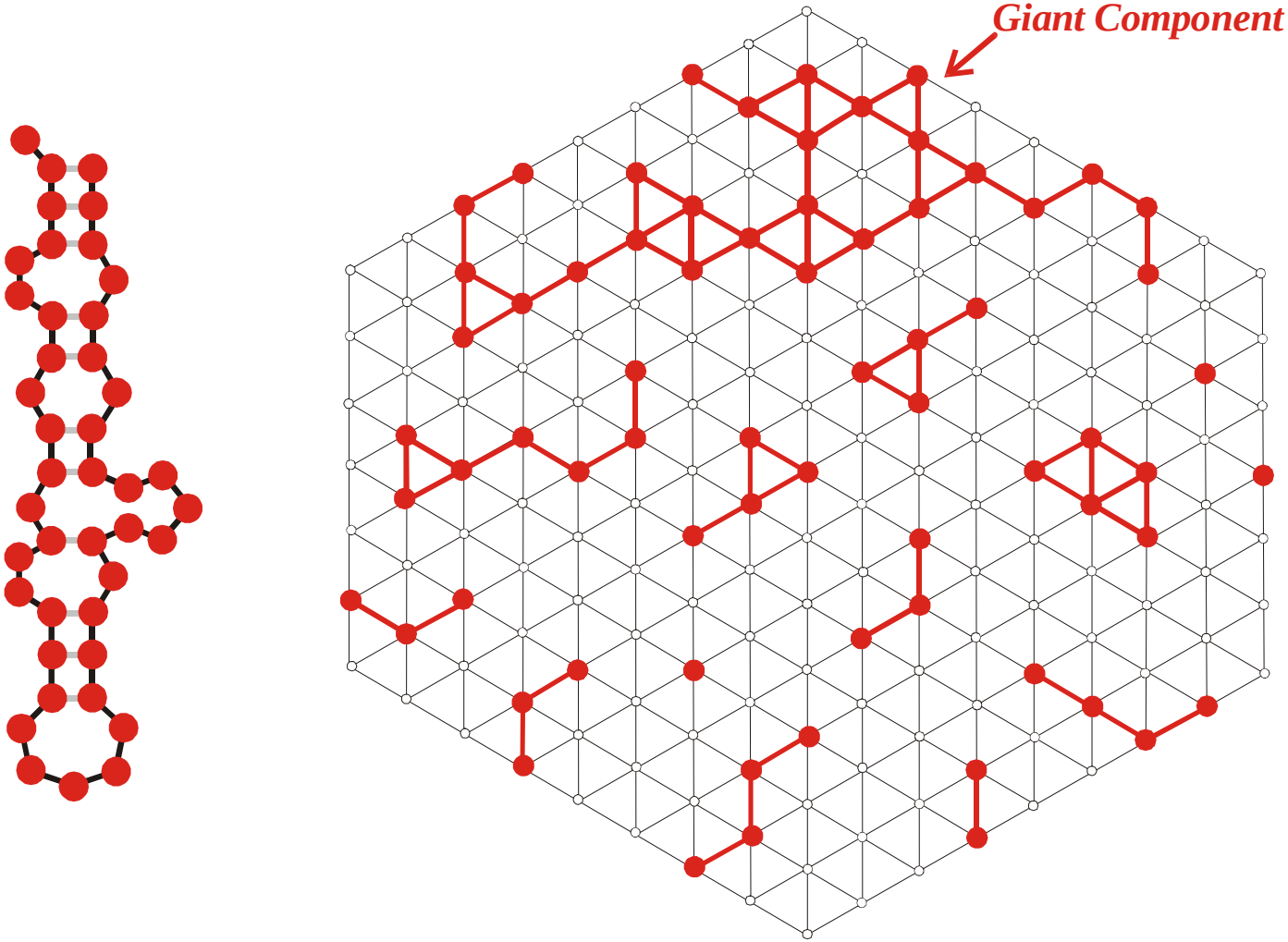
$\bar{\lambda}_k < \lambda_{cr}$ .... network G_k is **not** connected

κ	λ_{cr}	
2	0.5	**GC,AU**
3	0.423	**GUC,AUG**
4	0.370	**AUGC**

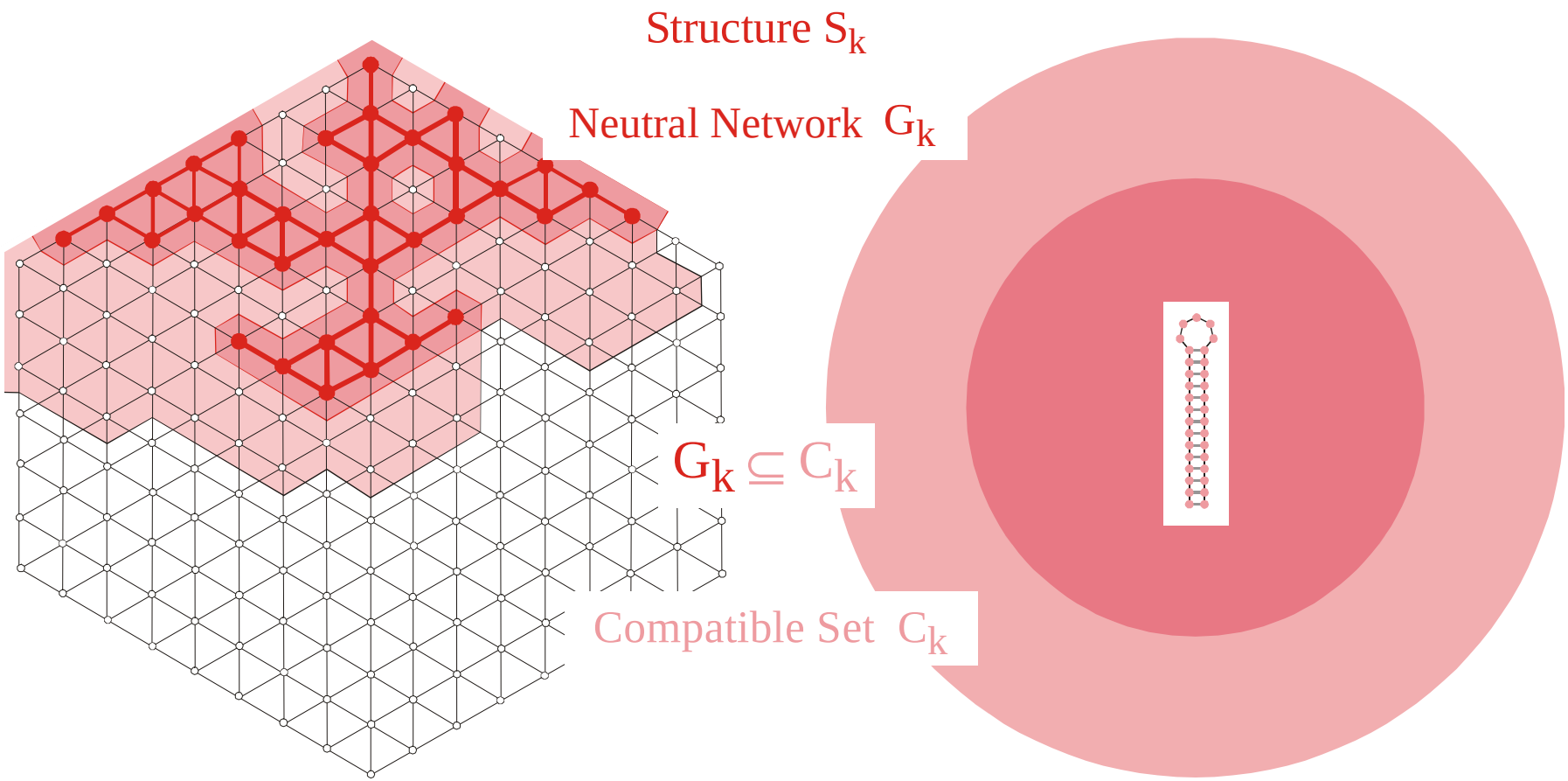
Mean degree of neutrality and connectivity of neutral networks



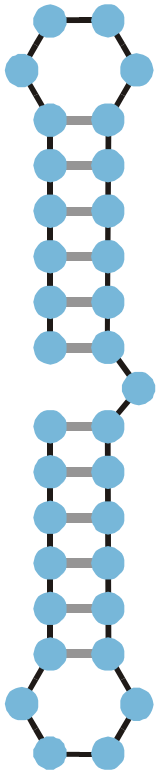
A connected neutral network formed by a common structure



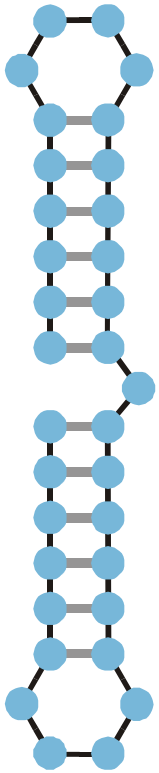
A multi-component neutral network formed by a rare structure



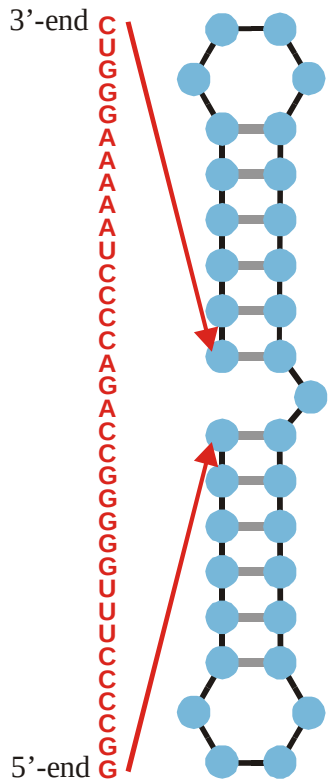
The **compatible set C_k** of a structure **S_k** consists of all sequences which form **S_k** as its minimum free energy structure (the neutral network **G_k**) or one of its suboptimal structures.



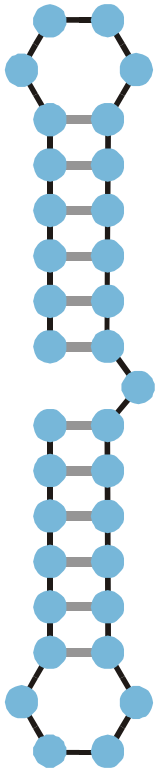
Structure



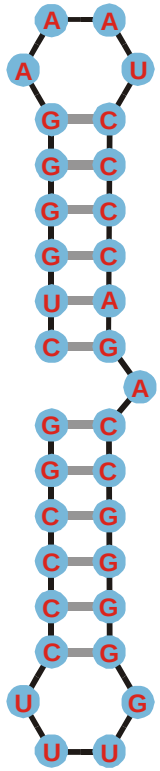
Structure



Compatible sequence

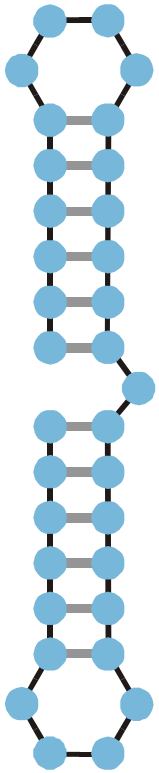


3’-end C U G G G A A A A A A U C C C C C A G A C C G G G G U U U C C C C G G 5’-end



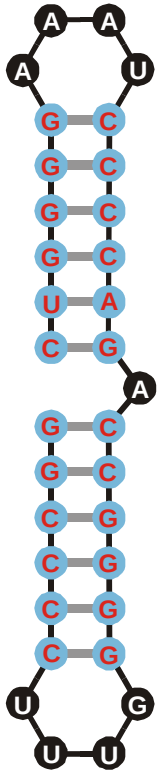
Structure

Compatible sequence



Structure

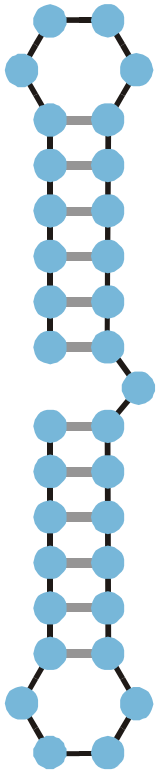
3’-end C U G G G A A A A A A U C C C C C A G A C C G G G G G U U U C C C C G G 5’-end



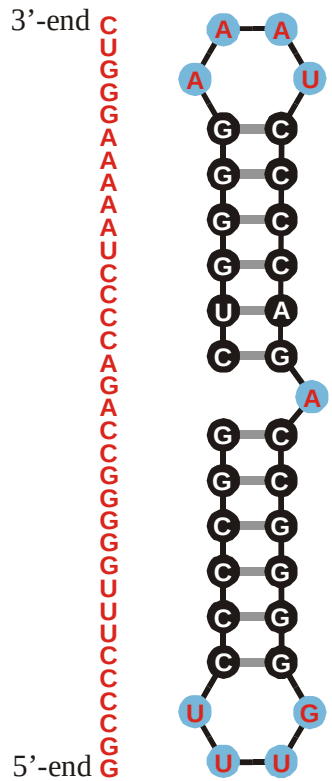
Compatible sequence

Single nucleotides: **A,U,G,C**

Single bases pairs are varied independently



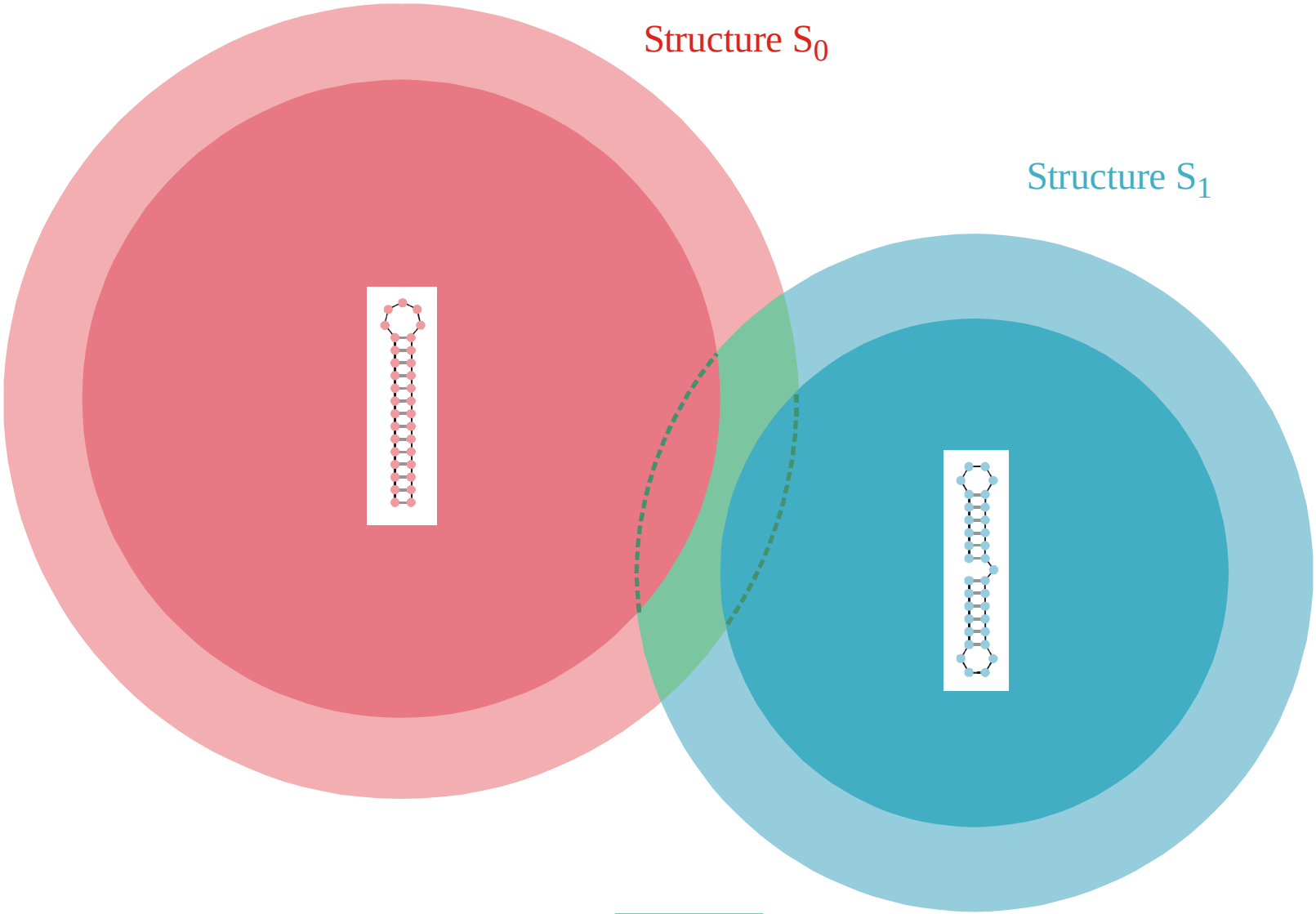
Structure



Compatible sequence

Base pairs: AU , UA
GC , CG
GU , UG

Base pairs are varied in strict correlation



Intersection of two compatible sets: $C_0 \cap C_1$

The intersection of two compatible sets is always non empty: $C_0 \cap C_1 \notin \varnothing$

Bulletin of Mathematical Biology, Vol. 59, No. 2, pp. 339–397, 1997
Elsevier Science Inc.

0092-8240/97 $17.00 + 0.00

S0092-8240(96)00089-4

GENERIC PROPERTIES OF COMBINATORY MAPS: NEUTRAL NETWORKS OF RNA SECONDARY STRUCTURES[1]

■ CHRISTIAN REIDYS*,†, PETER F. STADLER*,‡ and PETER SCHUSTER*,‡,§,[2]
*Santa Fe Institute,
Santa Fe, NM 87501, U.S.A.

†Los Alamos National Laboratory,
Los Alamos, NM 87545, U.S.A.

‡Institut für Theoretische Chemie der Universität Wien,
A-1090 Wien, Austria

§Institut für Molekulare Biotechnologie,
D-07708 Jena, Germany

(*E.mail: pks@tbi.univie.ac.at*)

Random graph theory is used to model and analyse the relationships between sequences and secondary structures of RNA molecules, which are understood as mappings from sequence space into shape space. These maps are non-invertible since there are always many orders of magnitude more sequences than structures. Sequences folding into identical structures form *neutral networks*. A neutral network is embedded in the set of sequences that are *compatible* with the given structure. Networks are modeled as graphs and constructed by random choice of vertices from the space of compatible sequences. The theory characterizes neutral networks by the mean fraction of neutral neighbors (λ). The networks are connected and percolate sequence space if the fraction of neutral nearest neighbors exceeds a threshold value ($\lambda > \lambda^*$). Below threshold ($\lambda < \lambda^*$), the networks are partitioned into a largest "giant" component and several smaller components. Structures are classified as "common" or "rare" according to the sizes of their pre-images, i.e. according to the fractions of sequences folding into them. The neutral networks of any pair of two different common structures almost touch each other, and, as expressed by the conjecture of *shape space covering* sequences folding into almost all common structures, can be found in a small ball of an arbitrary location in sequence space. The results from random graph theory are compared to data obtained by folding large samples of RNA sequences. Differences are explained in terms of specific features of RNA molecular structures. © *1997 Society for Mathematical Biology*

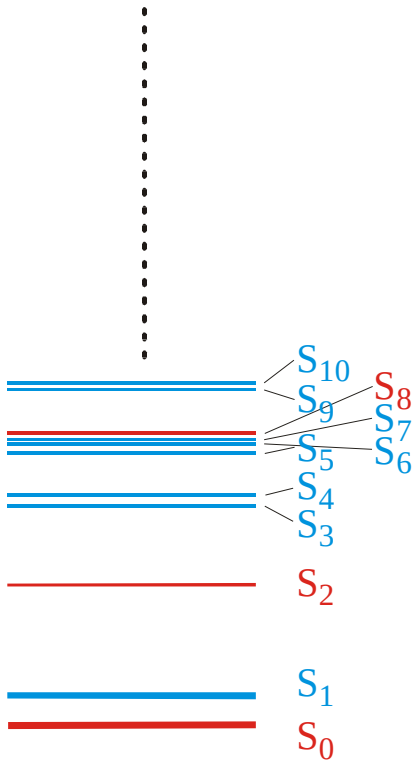
THEOREM 5. INTERSECTION-THEOREM. *Let* s *and* s′ *be arbitrary secondary structures and* **C**[s], **C**[s′] *their corresponding compatible sequences. Then,*

$$\mathbf{C}[s] \cap \mathbf{C}[s'] \neq \varnothing.$$

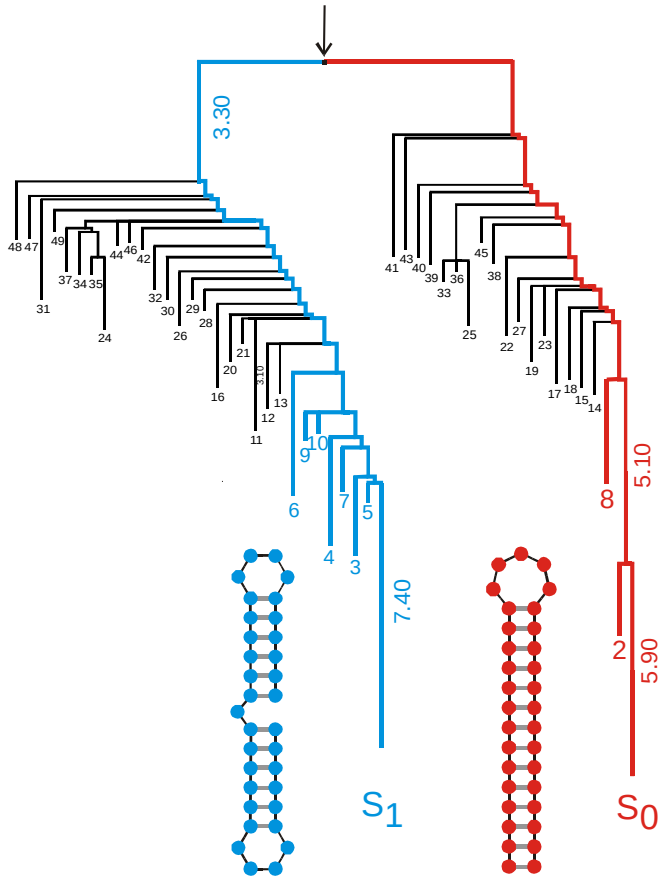
Proof. Suppose that the alphabet admits only the complementary base pair $[XY]$ and we ask for a sequence x compatible to both s and s'. Then $\gamma(s, s') \cong D_m$ operates on the set of all positions $\{x_1, \ldots, x_n\}$. Since we have the operation of a dihedral group, the orbits are either cycles or chains and the cycles have even order. A constraint for the sequence compatible to both structures appears only in the cycles where the choice of bases is not independent. It remains to be shown that there is a valid choice of bases for each cycle, which is obvious since these have even order. Therefore, it suffices to choose an alternating sequence of the pairing partners X and Y. Thus, there are at least two different choices for the first base in the orbit. ■

Remark. A generalization of the statement of theorem 5 to three different structures is false.

Reference for the definition of the intersection and the proof of the **intersection theorem**

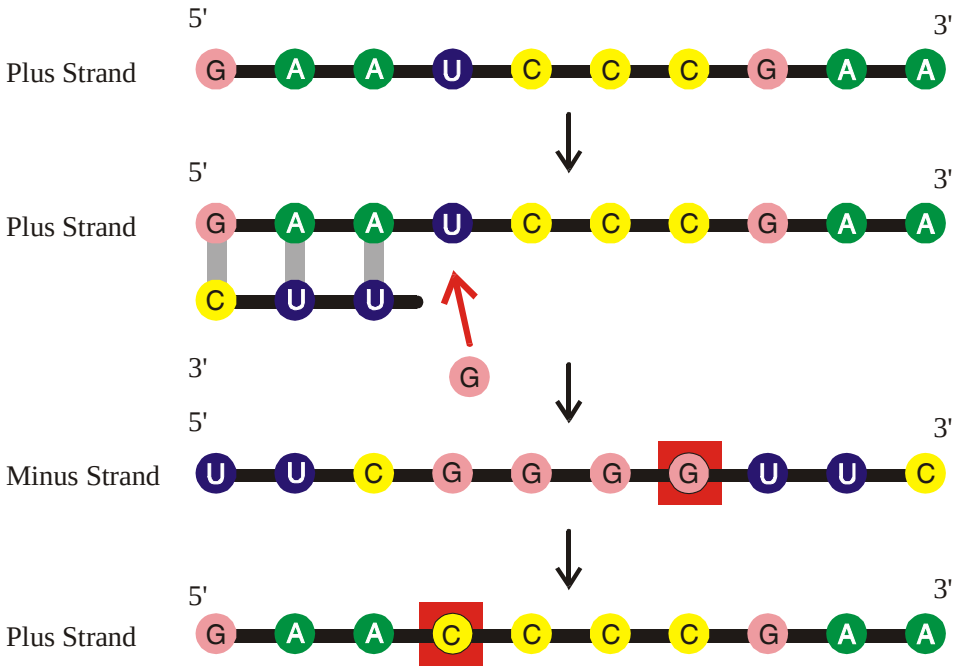


Suboptimal structures



Kinetic folding

A typical energy landscape of a sequence with two (meta)stable comformations



Point Mutation

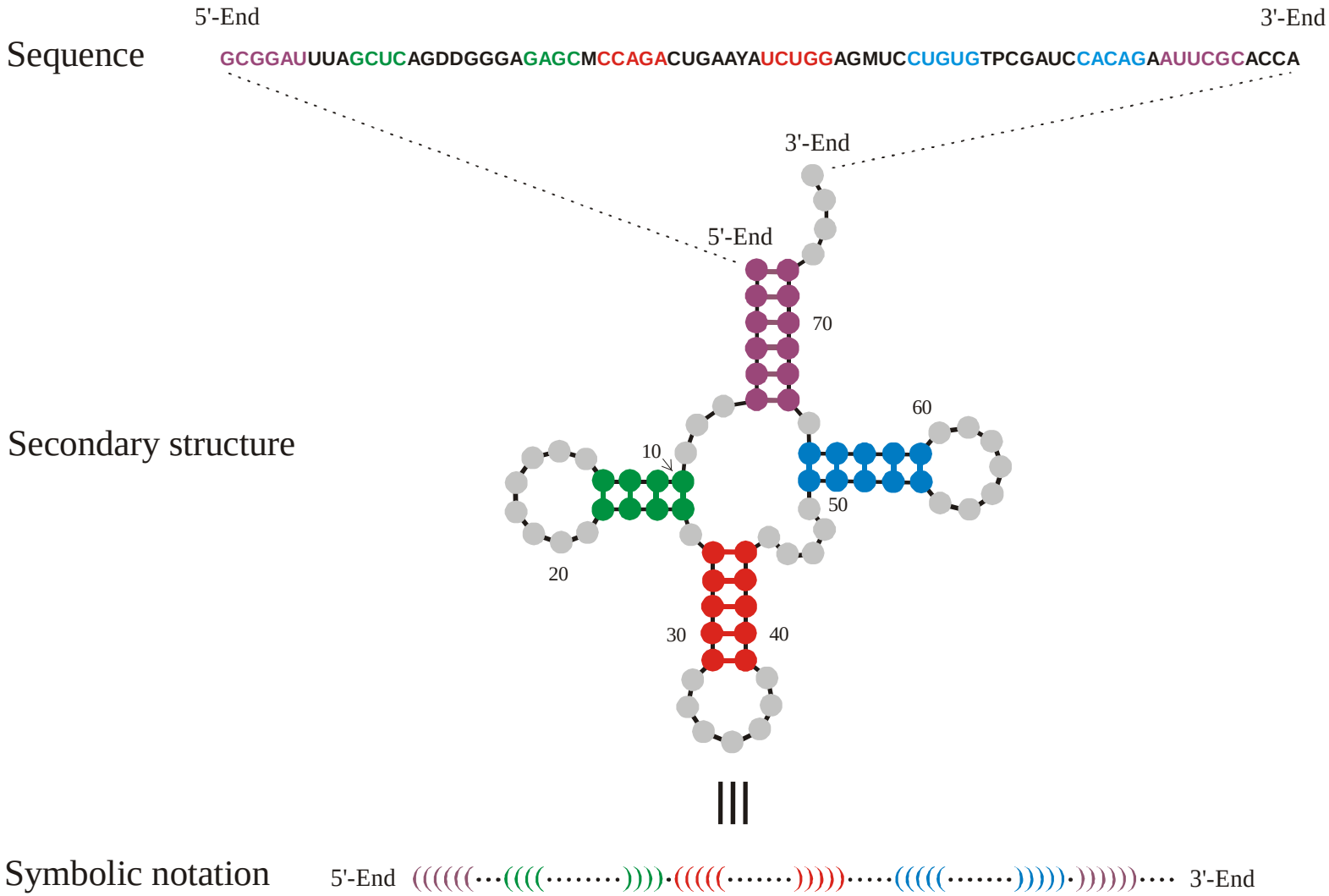
GAAUCCCGAA → GAAUCCCGUCCCGAA

Insertion

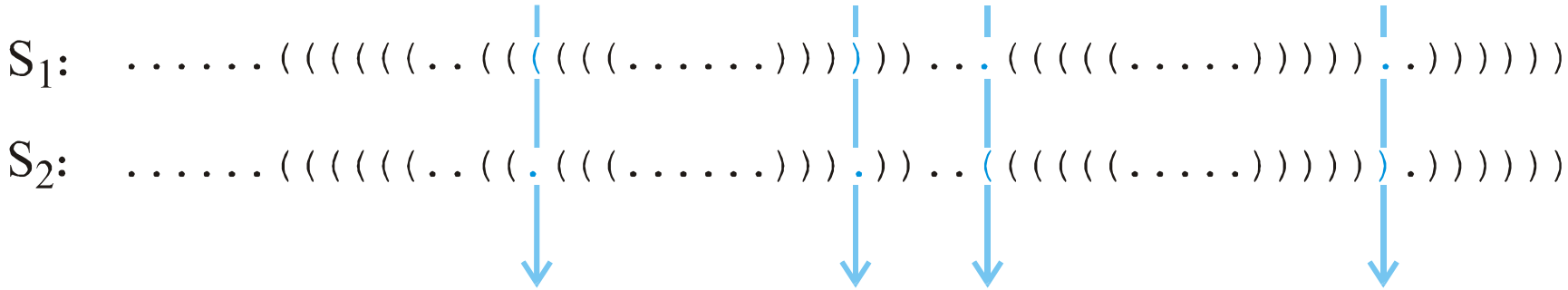
GAAUCCCGAA → GAAUCCA

Deletion

The origins of changes in RNA sequences are **replication errors** called **mutations**.



A symbolic notation of RNA secondary structure that is equivalent to the conventional graphs



Hamming distance $d_H(S_1,S_2) = 4$

(i) $d_H(S_1,S_1) = 0$

(ii) $d_H(S_1,S_2) = d_H(S_2,S_1)$

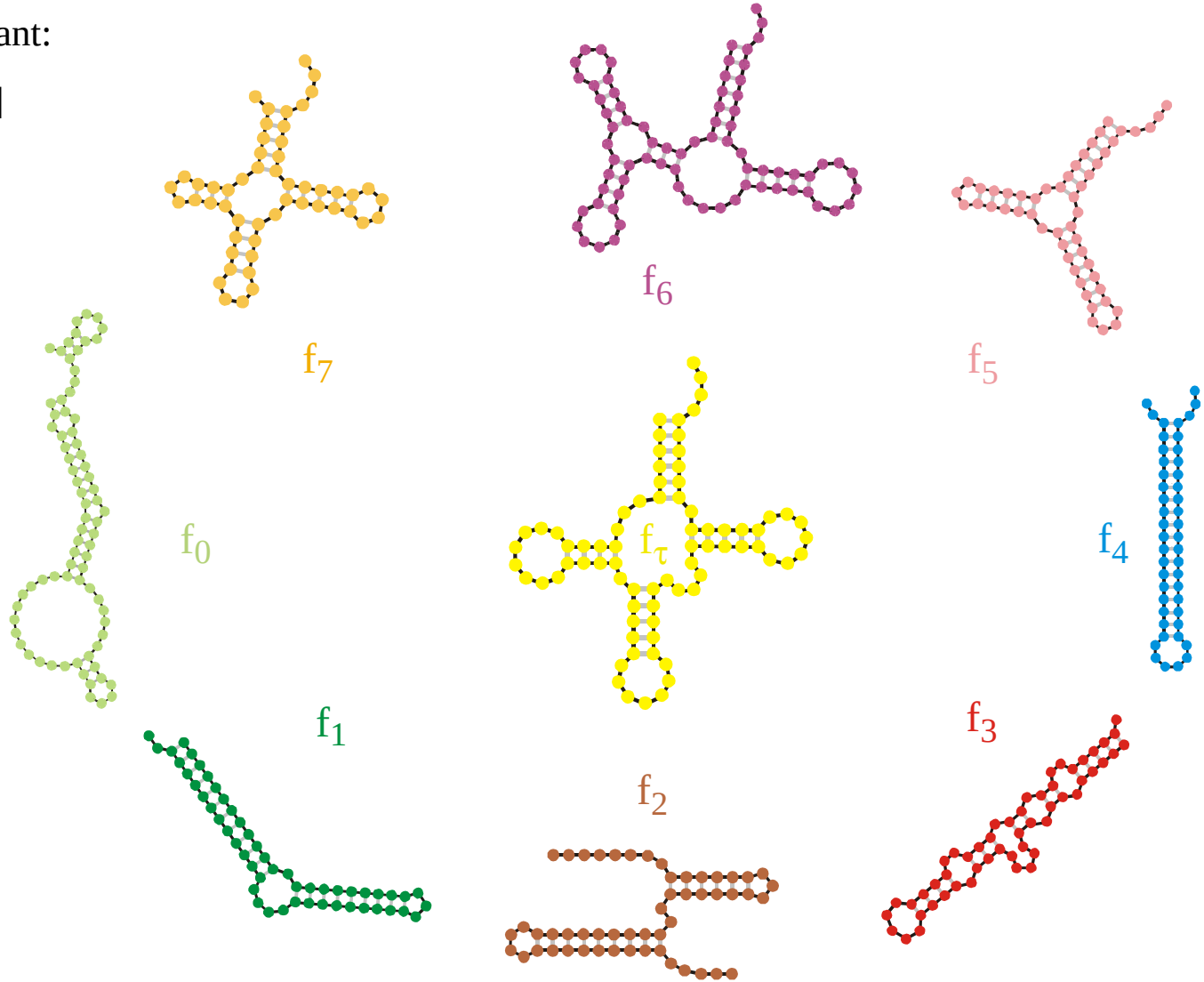
(iii) $d_H(S_1,S_3) \leq d_H(S_1,S_2) + d_H(S_2,S_3)$

The Hamming distance between structures in parentheses notation forms a metric in structure space

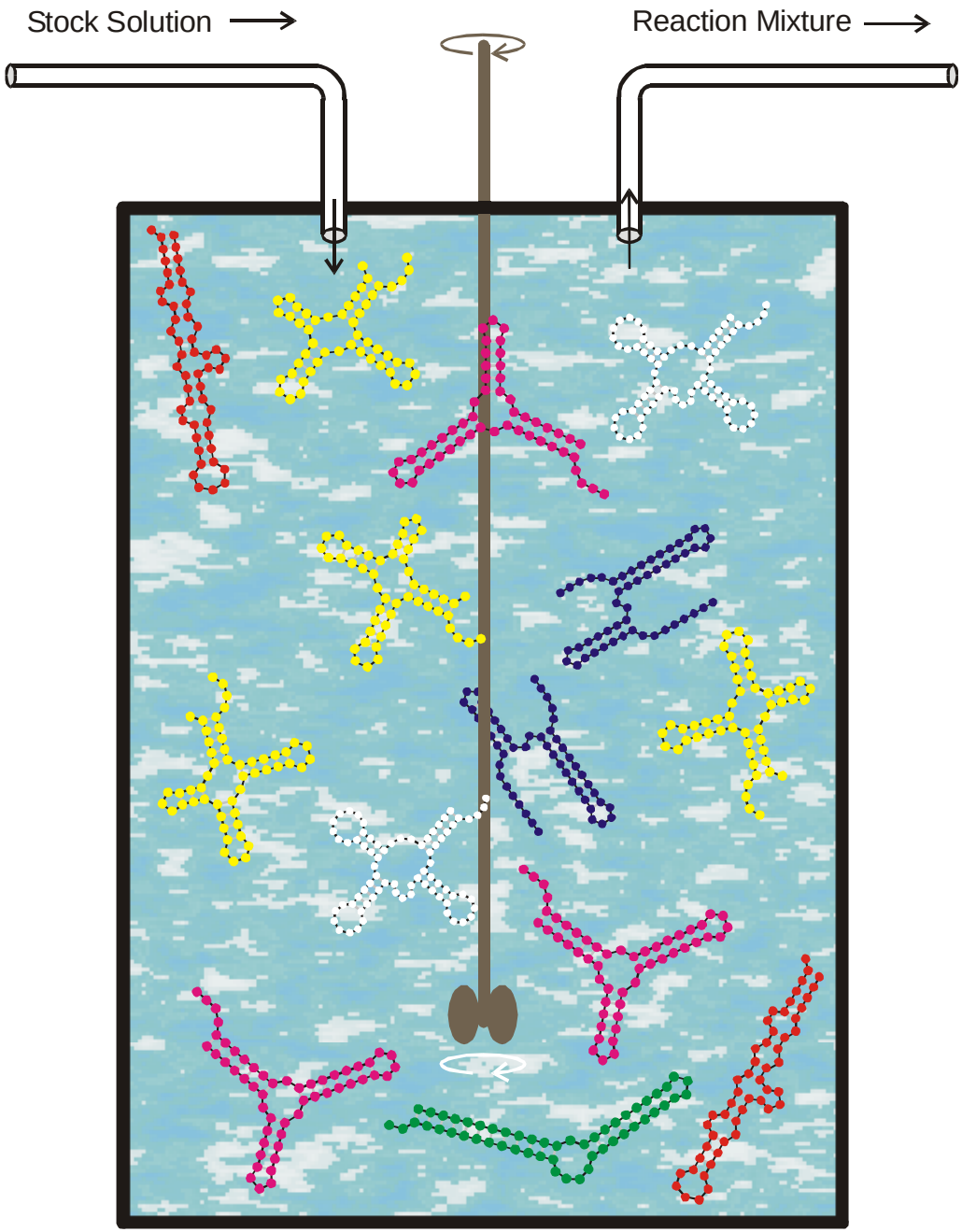
Replication rate constant:

$$f_k = \gamma / [\alpha + \Delta d_S^{(k)}]$$

$$\Delta d_S^{(k)} = d_H(S_k, S_\tau)$$



Evaluation of RNA secondary structures yields replication rate constants



Replication rate constant:

$$f_k = \gamma / [\alpha + \Delta d_S^{(k)}]$$

$$\Delta d_S^{(k)} = d_H(S_k, S_\tau)$$

Selection constraint:

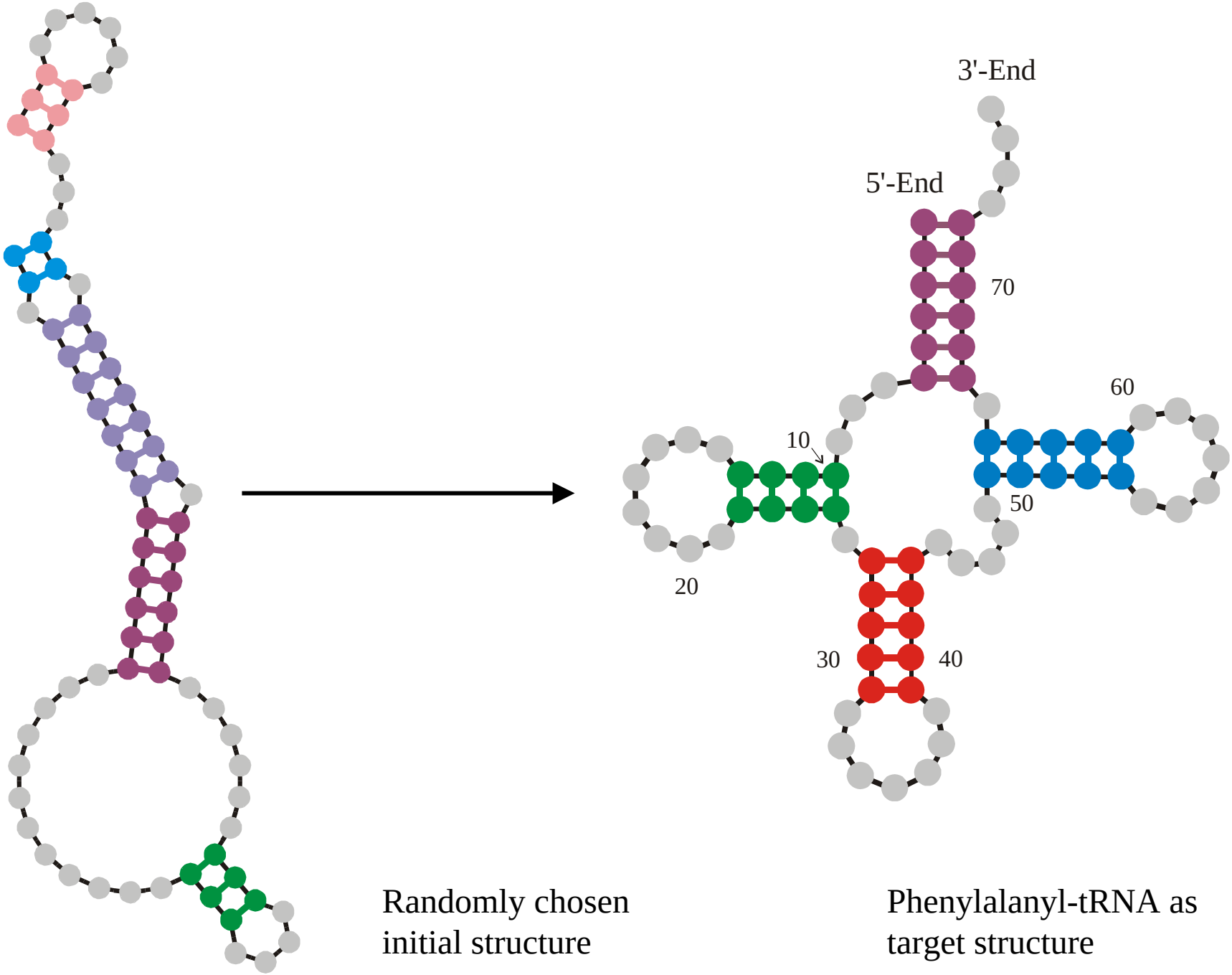
Population size, N = # RNA molecules, is controlled by the flow

$$N(t) \approx \overline{N} \pm \sqrt{\overline{N}}$$

Mutation rate:

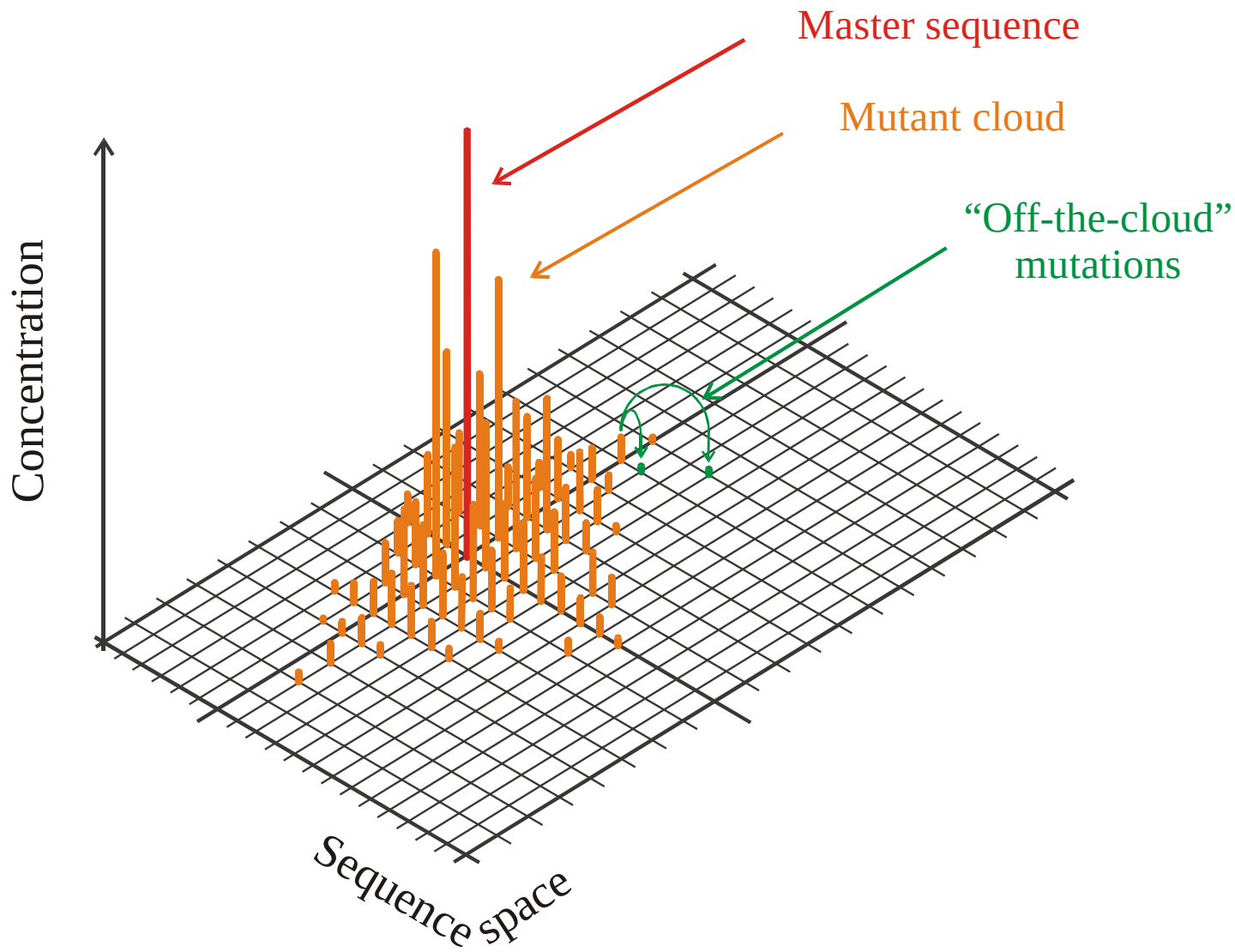
p = 0.001 / site × replication

The flowreactor as a device for **studies** of evolution ***in vitro*** and ***in silico***

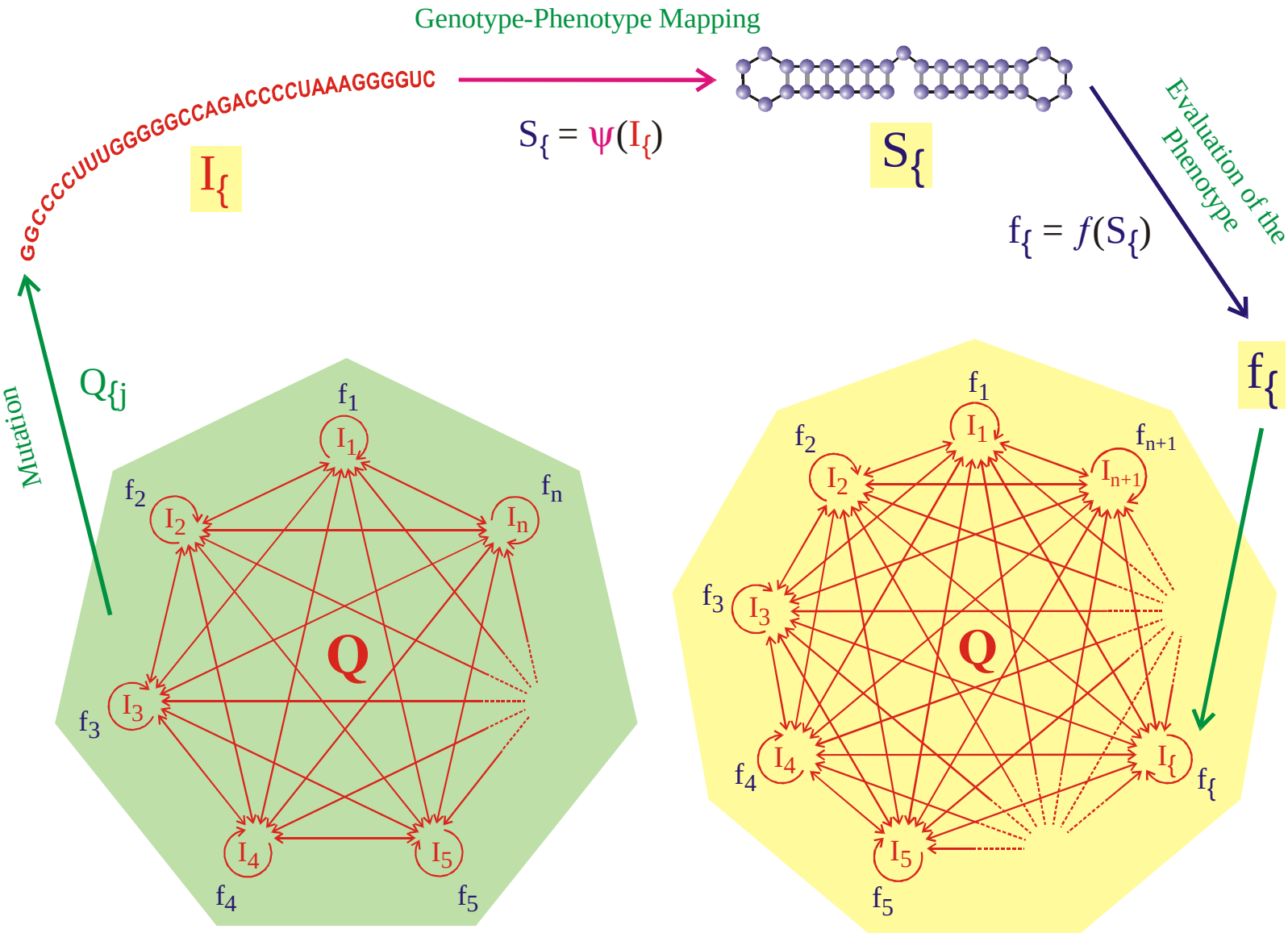


Randomly chosen initial structure

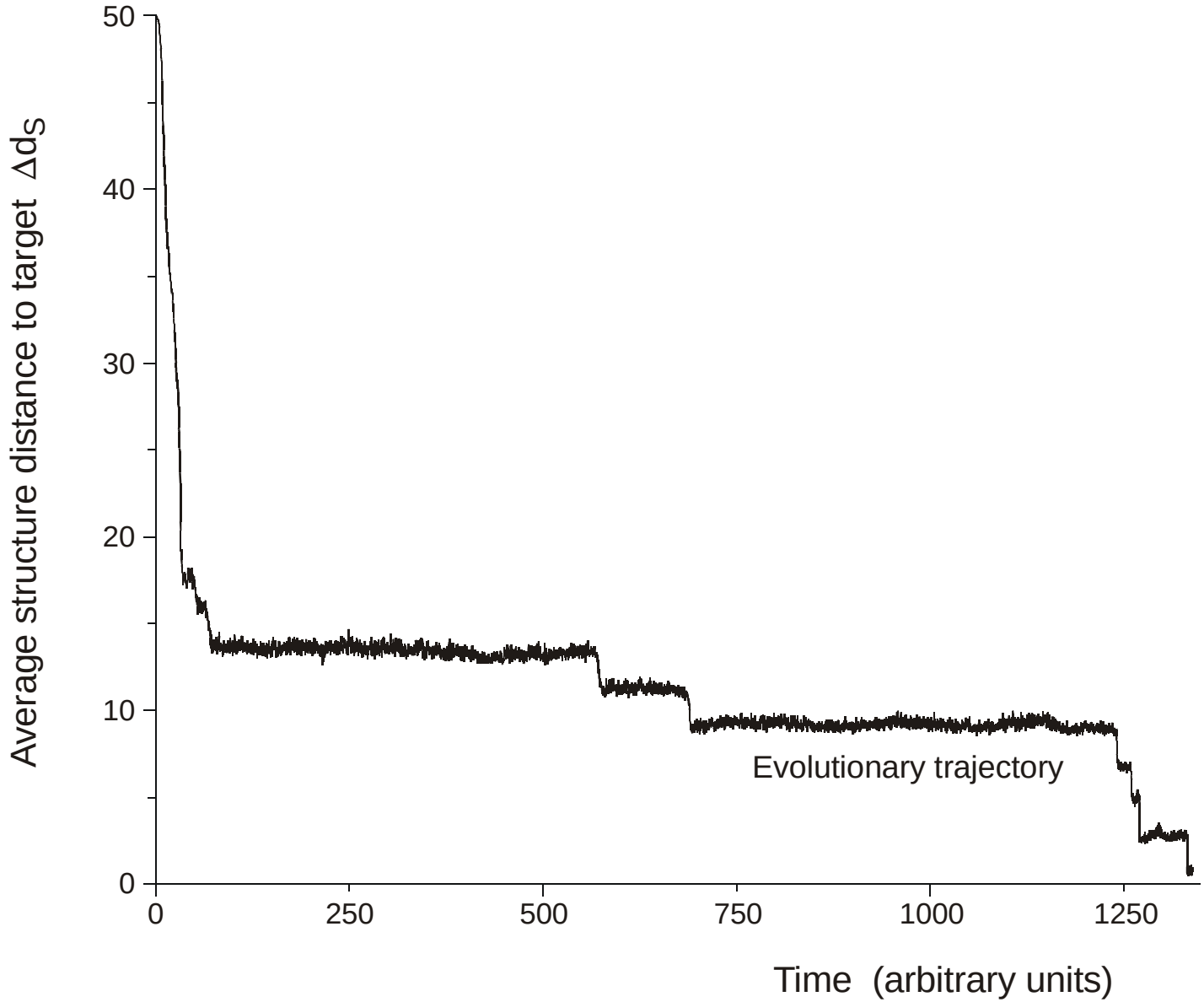
Phenylalanyl-tRNA as target structure



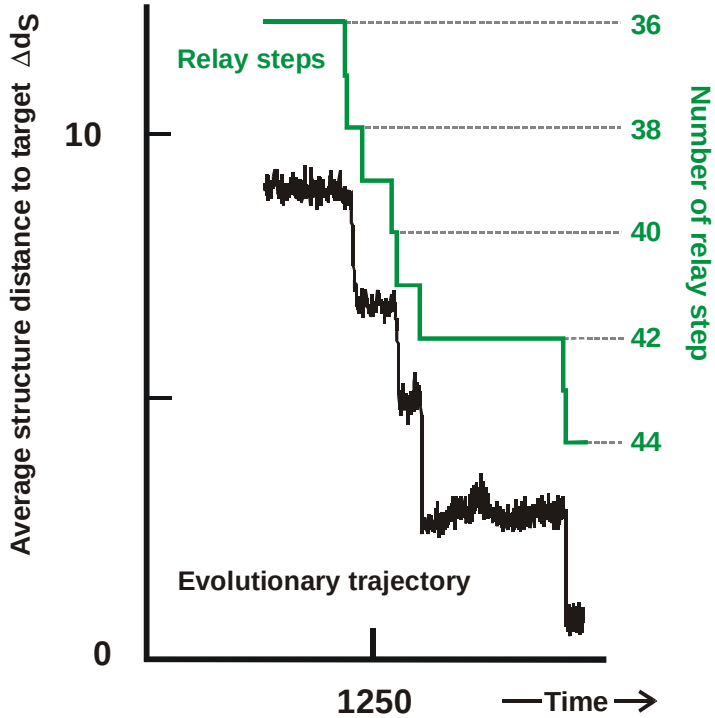
The molecular quasispecies
in sequence space

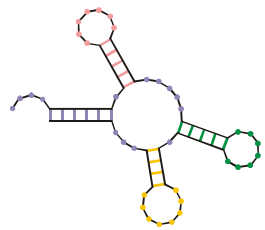


Evolutionary dynamics
including molecular phenotypes



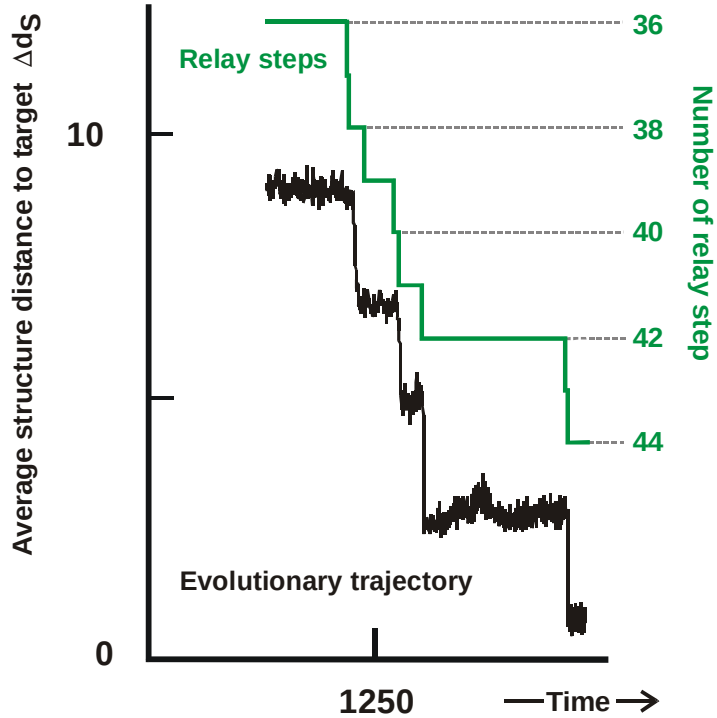
In silico optimization in the flow reactor: Evolutionary trajectory

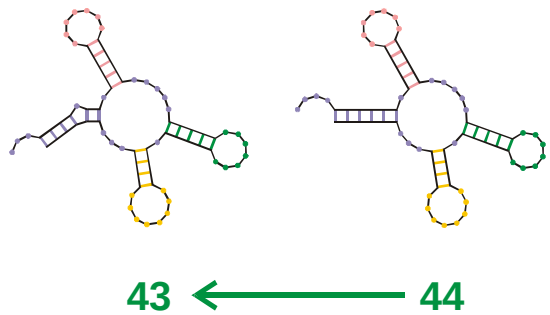




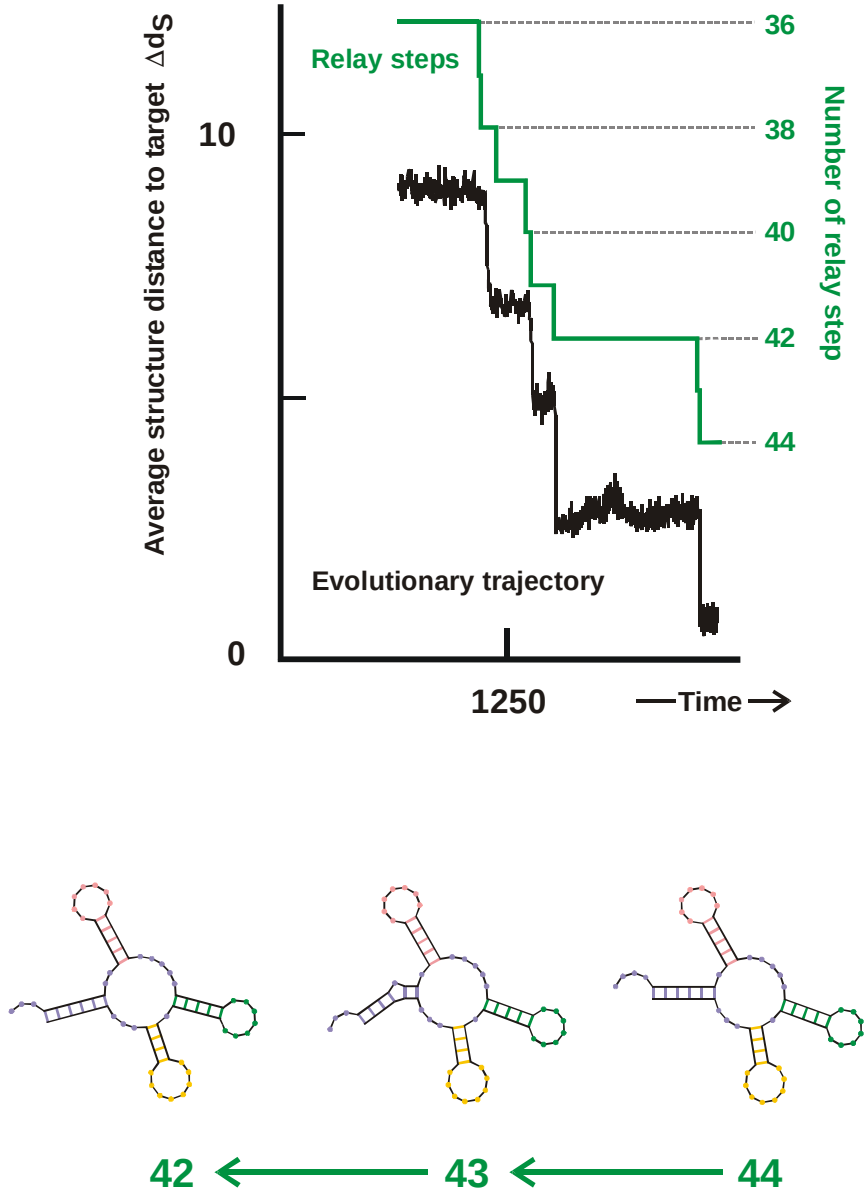
44

Final conformation of optimization

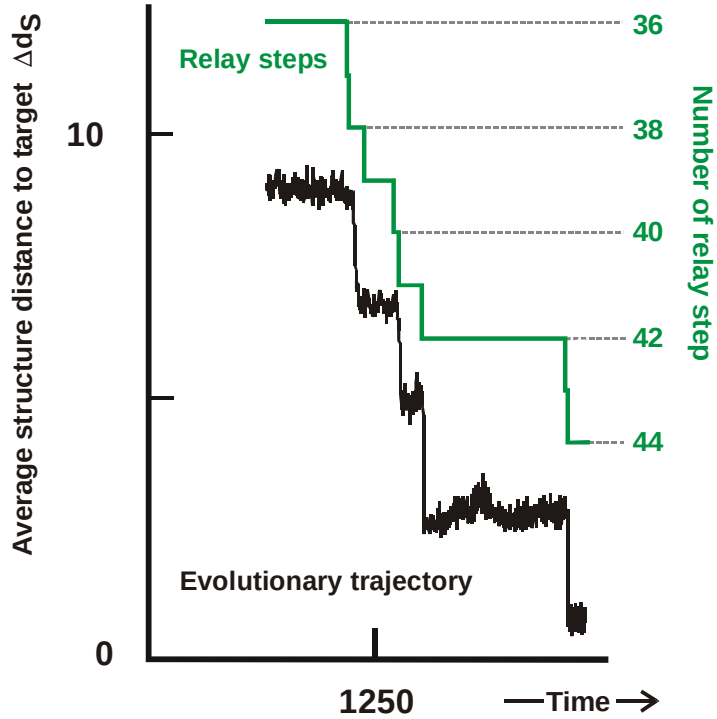


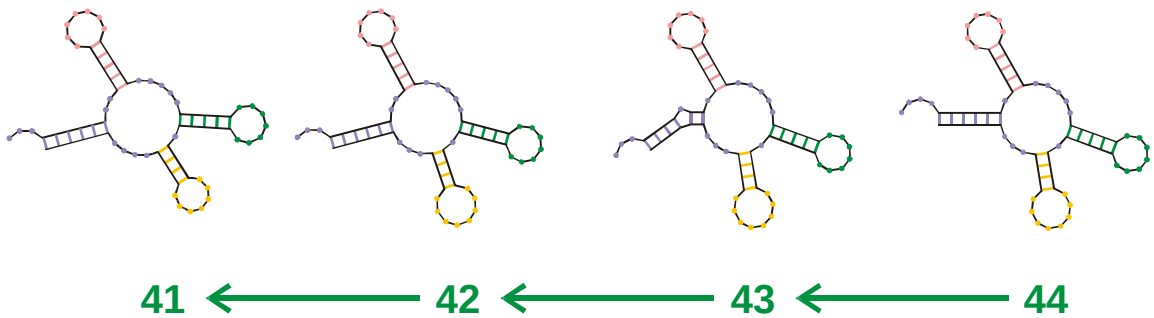


Reconstruction of the last step 43 → 44

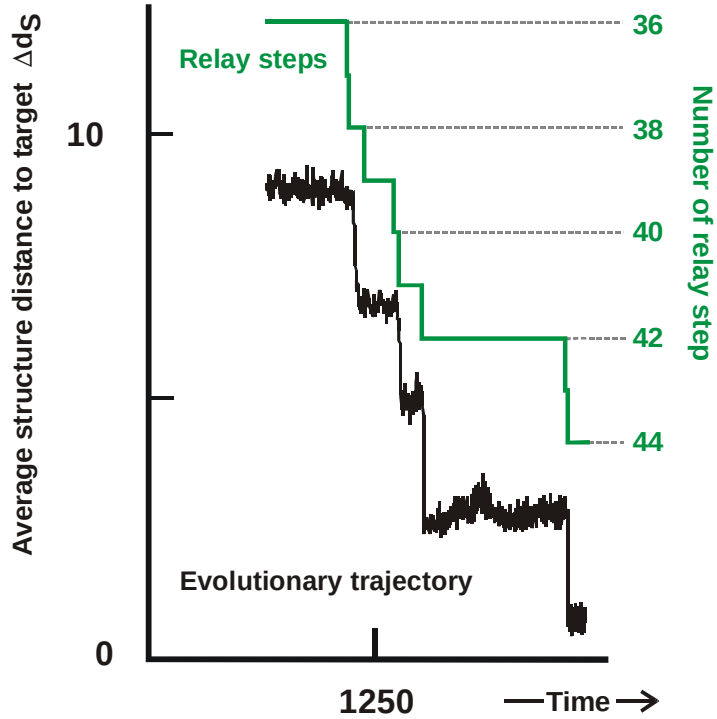


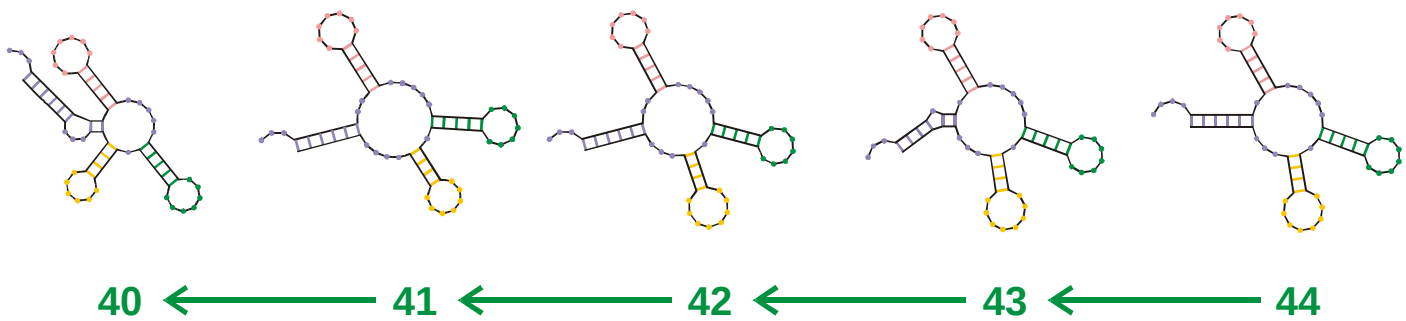
Reconstruction of last-but-one step 42 → 43 (→ 44)



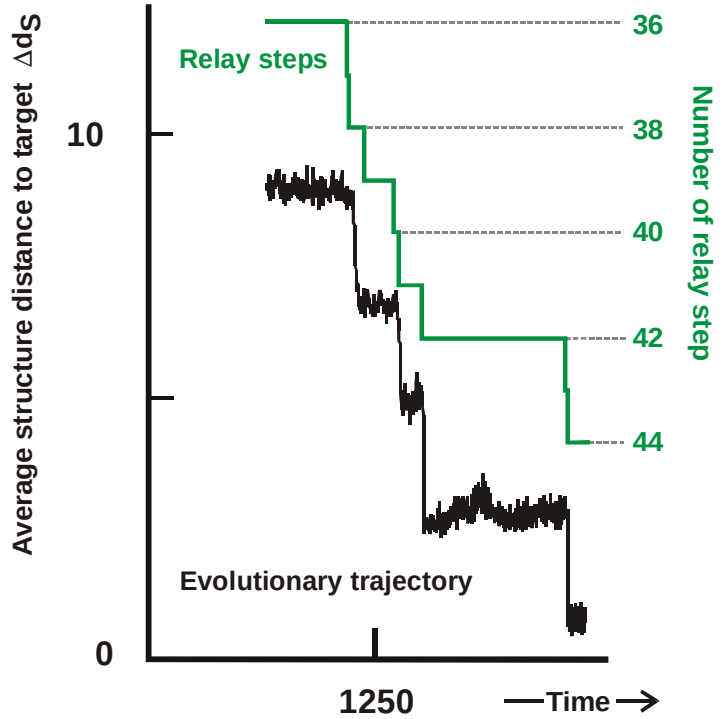


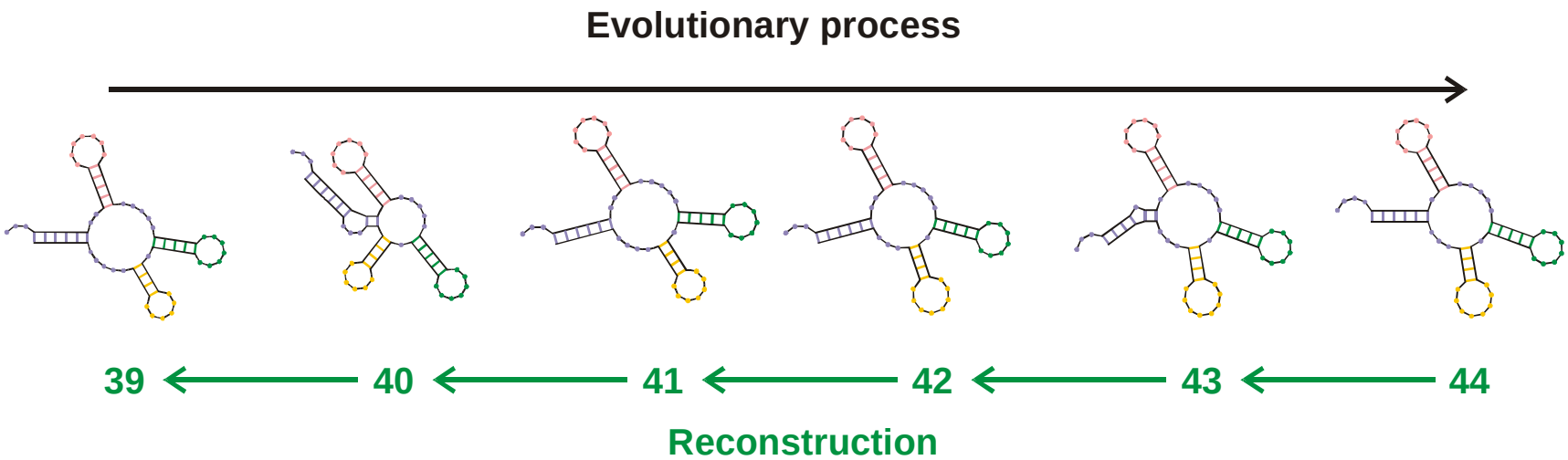
Reconstruction of step 41 → 42 (→ 43 → 44)



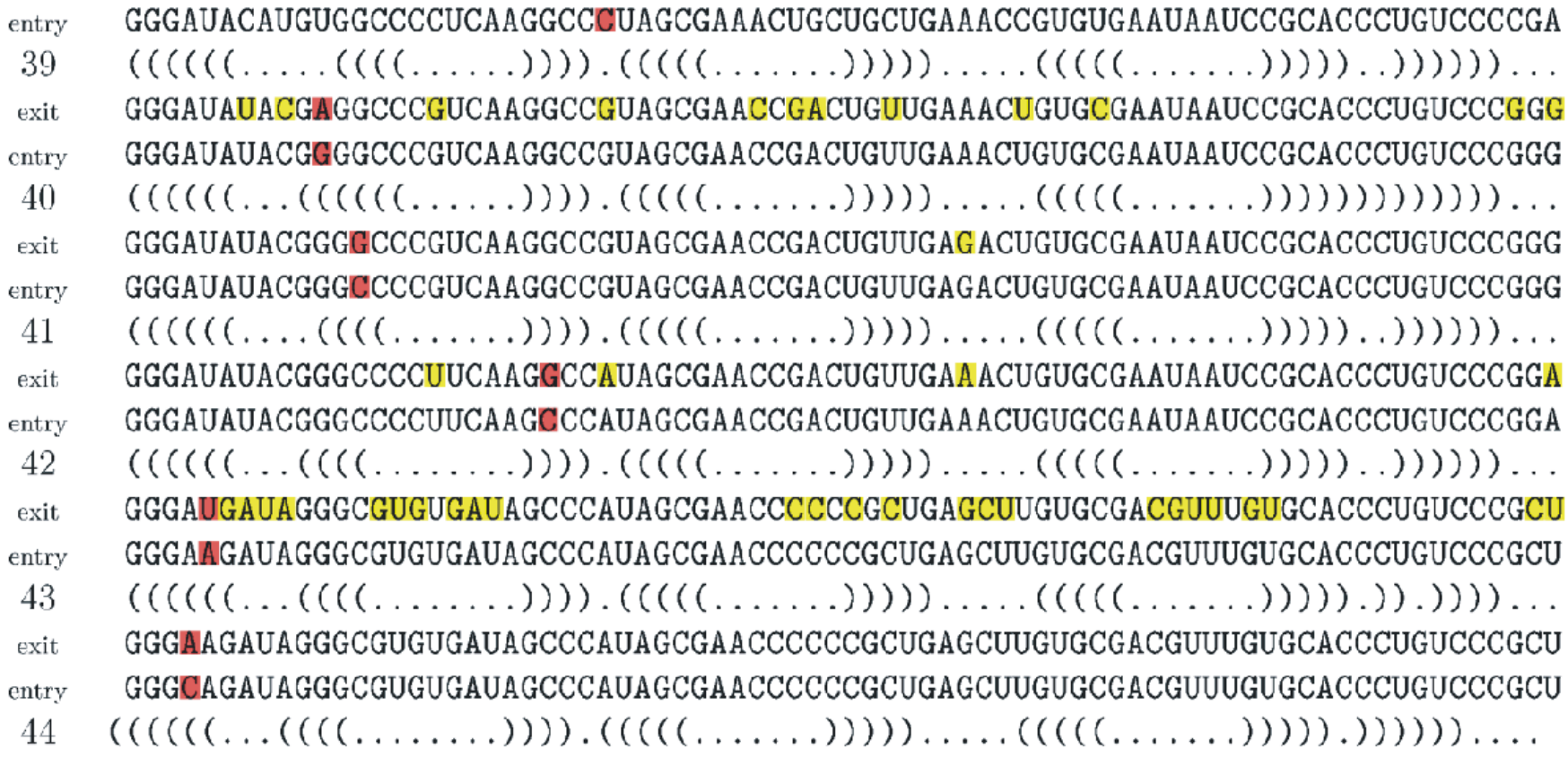


Reconstruction of step 40 → 41 (→ 42 → 43 → 44)





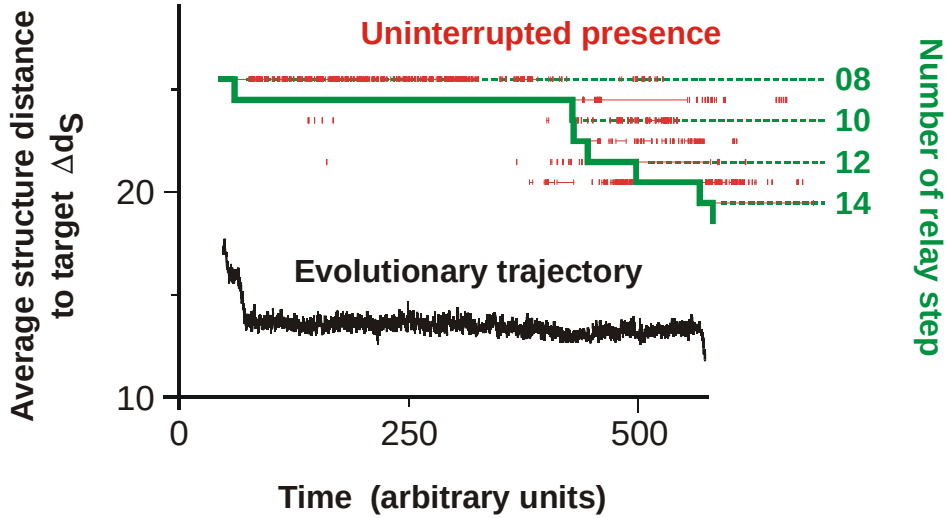
Reconstruction of the relay series



Transition inducing point mutations **Neutral point mutations**

Change in RNA sequences during the final five relay steps 39 → 44

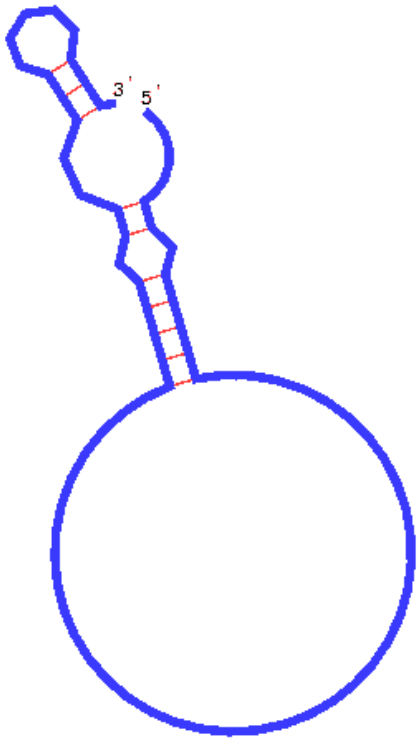
28 neutral point mutations during a long quasi-stationary epoch



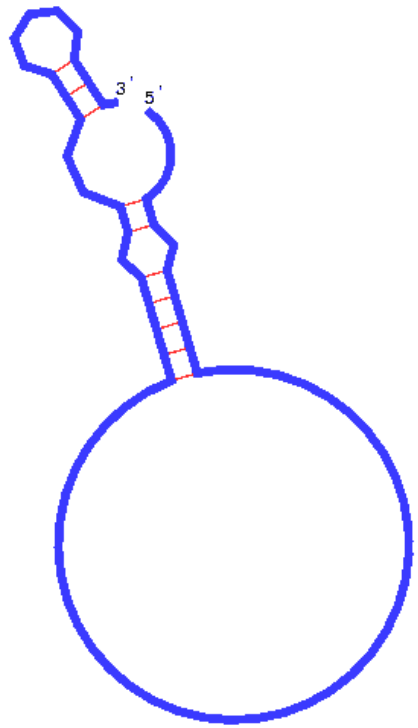

```
entry   GGUAUGGGCGUUGAAUAGUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCGUACAGAA
8       .((((((((((((.........(((....)))........)))))).....((((.........))))))))))....
exit    GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUCCCAUACAGAA
entry   GGUAUGGGCGUUGAAUAAUAGGGUUUAAACCAAUCGGCCAACGAUCUCGUGUGCGCAUUUCAUAUACCAUACAGAA
9       .((((((.((((.........(((....)))........)))))).....((((.........))))).))))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACACCGUCCCAAG
entry   UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
10      .(((((..((((.........(((....)))........)))))).....((((.........))))).)))))....
exit    UGGAUGGACGUUGAAUAACAAGGUAUCGACCAAACAACCAACGAGUAAGUGUGUACGCCCCACACAGCGUCCCAAG
```

Transition inducing point mutations — **Neutral point mutations**

Neutral genotype evolution during phenotypic stasis

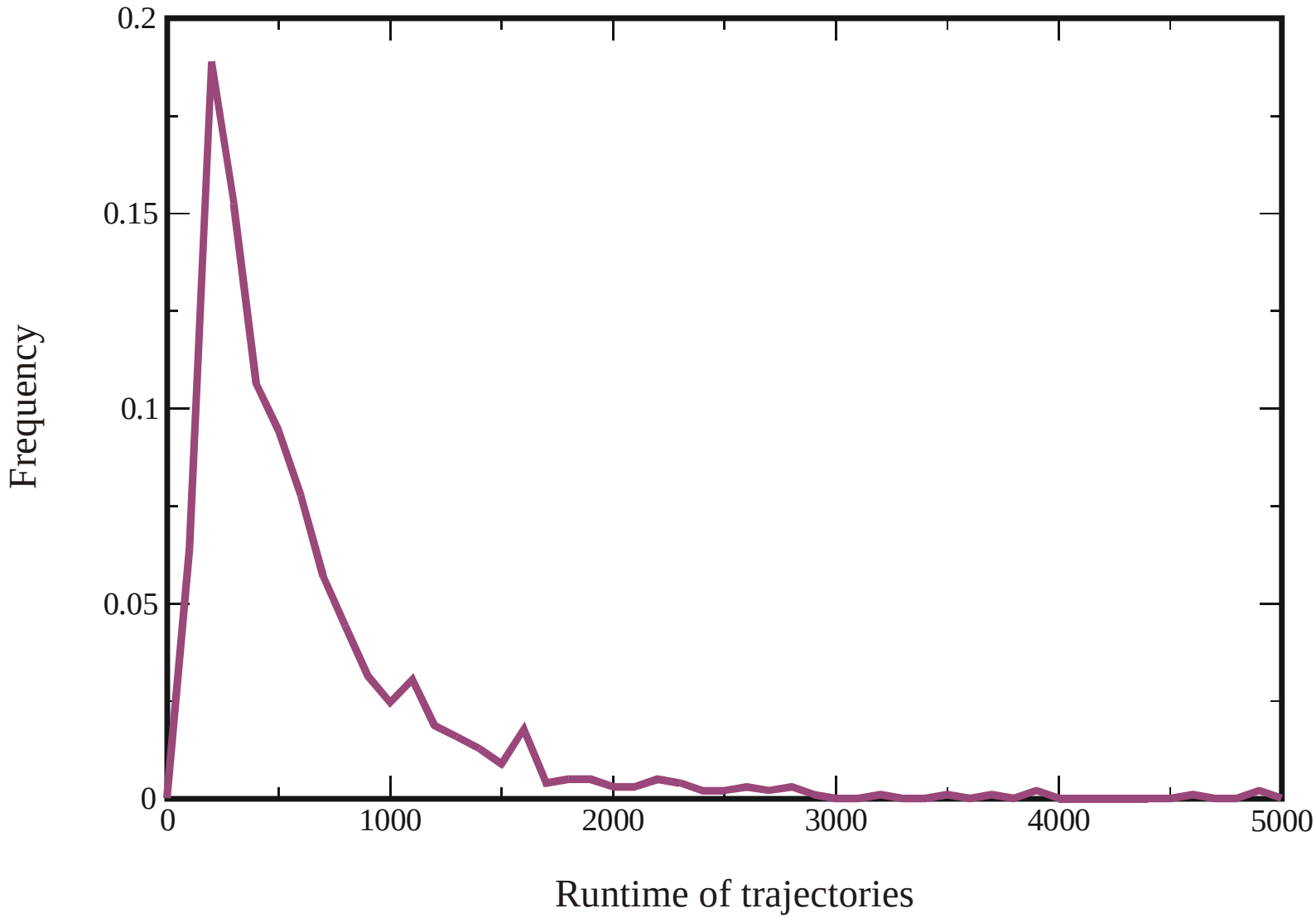


AUGC



GC

Movies of optimization trajectories over the **AUGC** and the **GC** alphabet



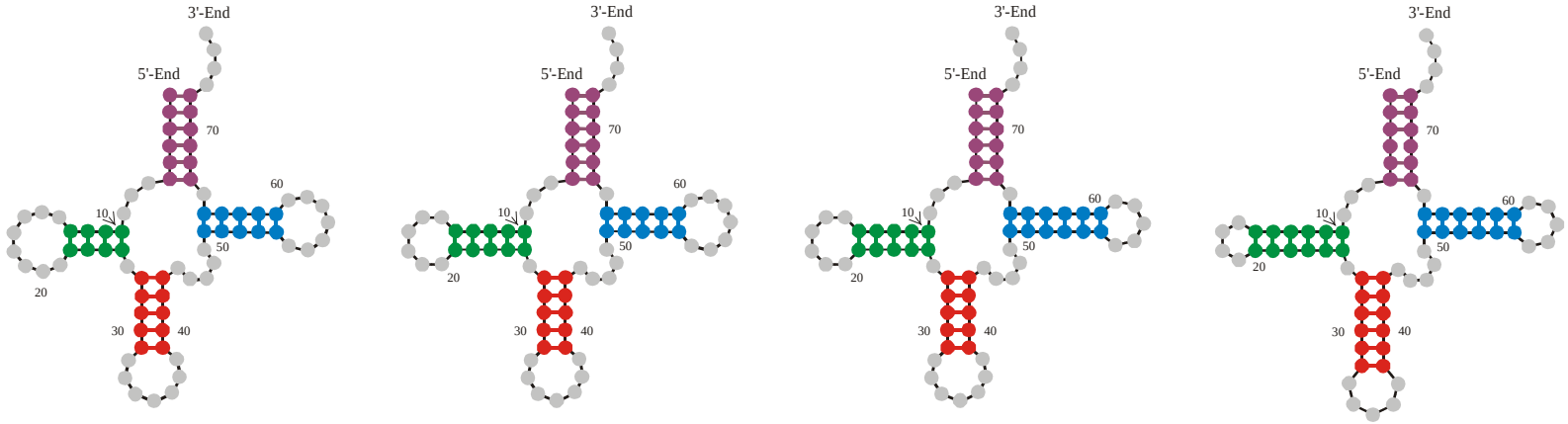
Statistics of the lengths of trajectories from initial structure to target (**AUGC**-sequences)

Alphabet	Runtime	Transitions	Main transitions	No. of runs
AUGC	385.6	22.5	12.6	1017
GUC	448.9	30.5	16.5	611
GC	2188.3	40.0	20.6	107

Mean population size: $N = 3000$; mutation rate: $p = 0.001$

Statistics of trajectories and relay series (mean values of log-normal distributions).

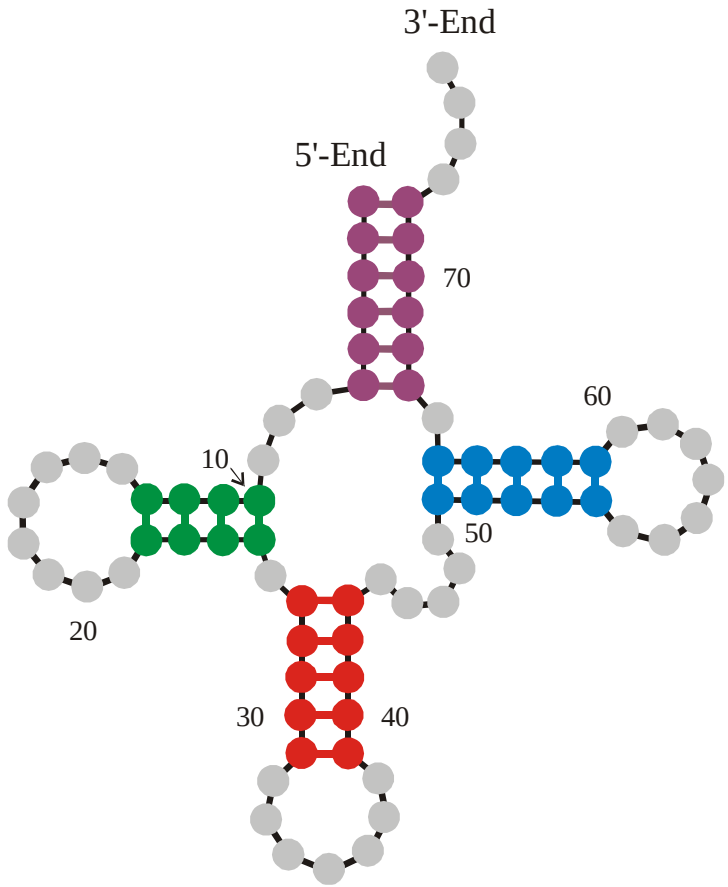
AUGC neutral networks of tRNAs are near the connectivity threshold, **GC** neutral networks are way below.



Alphabet	Degree of neutrality $\bar{\lambda}$			
AU	- -	- -	- -	0.073 ± 0.032
AUG	- -	0.217 ± 0.051	0.207 ± 0.055	0.201 ± 0.056
AUGC	0.275 ± 0.064	0.279 ± 0.063	0.289 ± 0.062	0.313 ± 0.058
UGC	0.263 ± 0.071	0.257 ± 0.070	0.251 ± 0.068	0.250 ± 0.064
GC	0.052 ± 0.033	0.057 ± 0.034	0.060 ± 0.033	0.068 ± 0.034

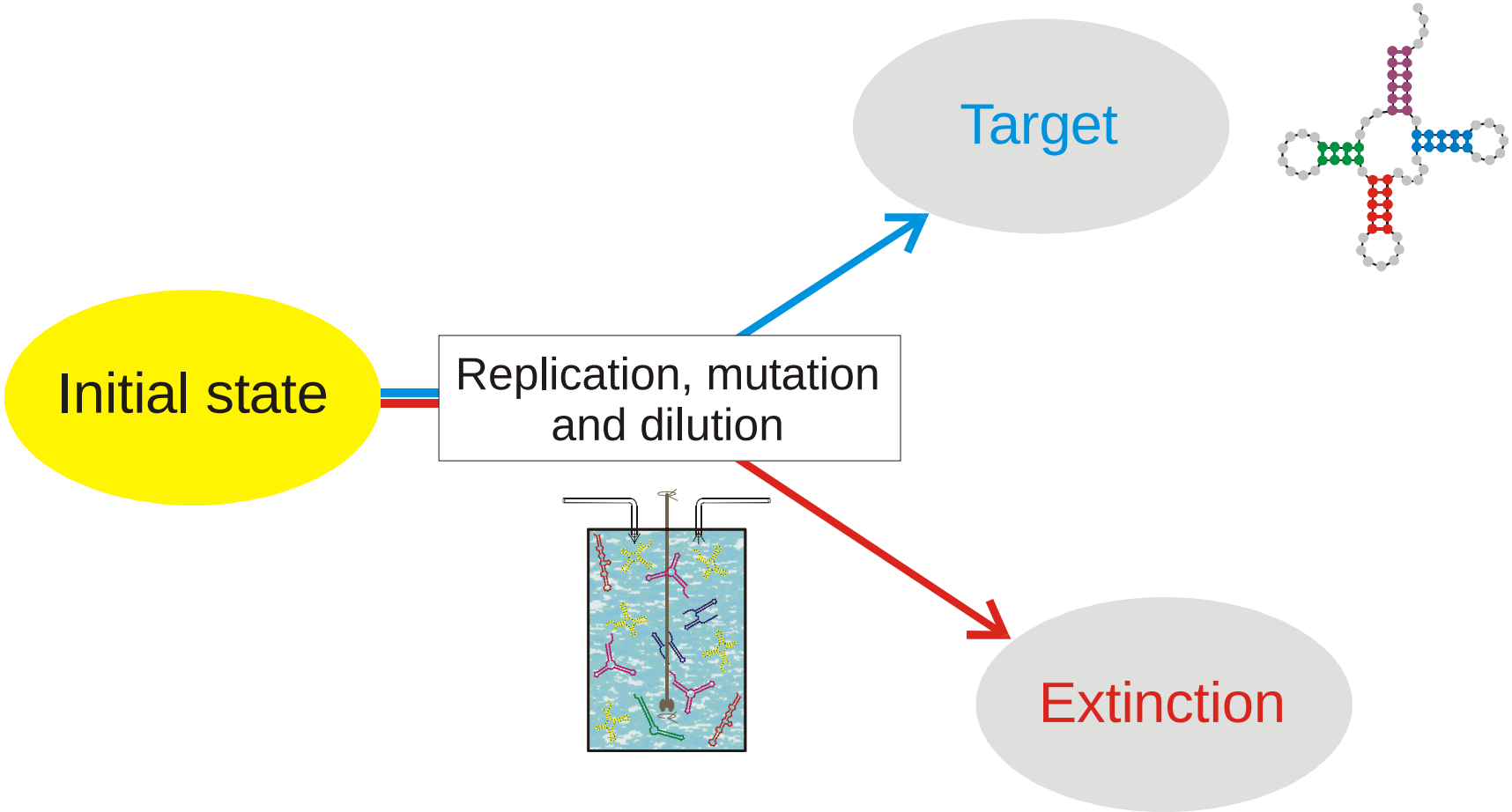
Degree of neutrality of cloverleaf RNA secondary structures over different alphabets

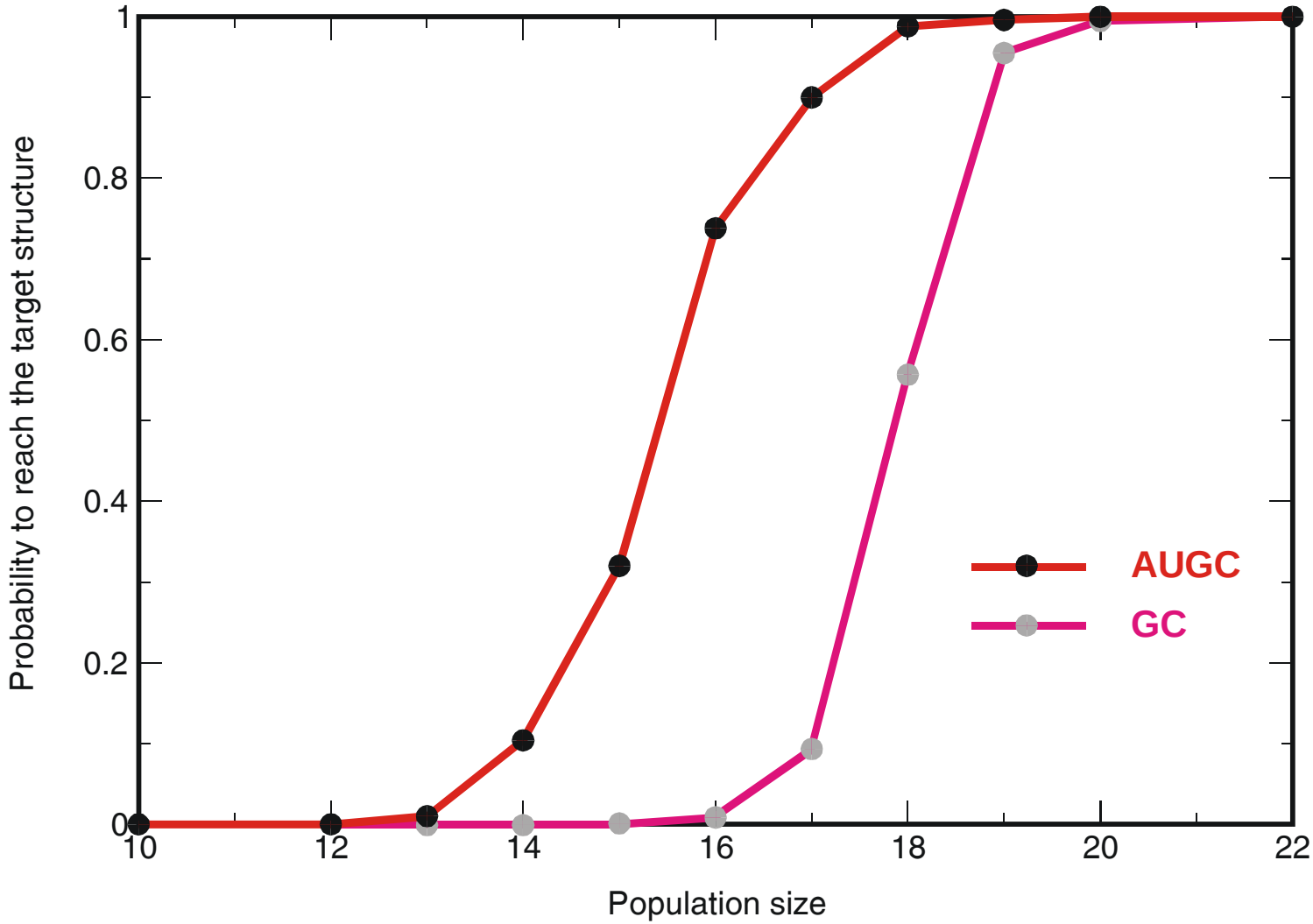
Stable tRNA clover leaf structures built from binary, **GC**-only, sequences exist. The corresponding sequences are found through inverse folding. Optimization by mutation and selection in the flow reactor turned out to be a hard problem.



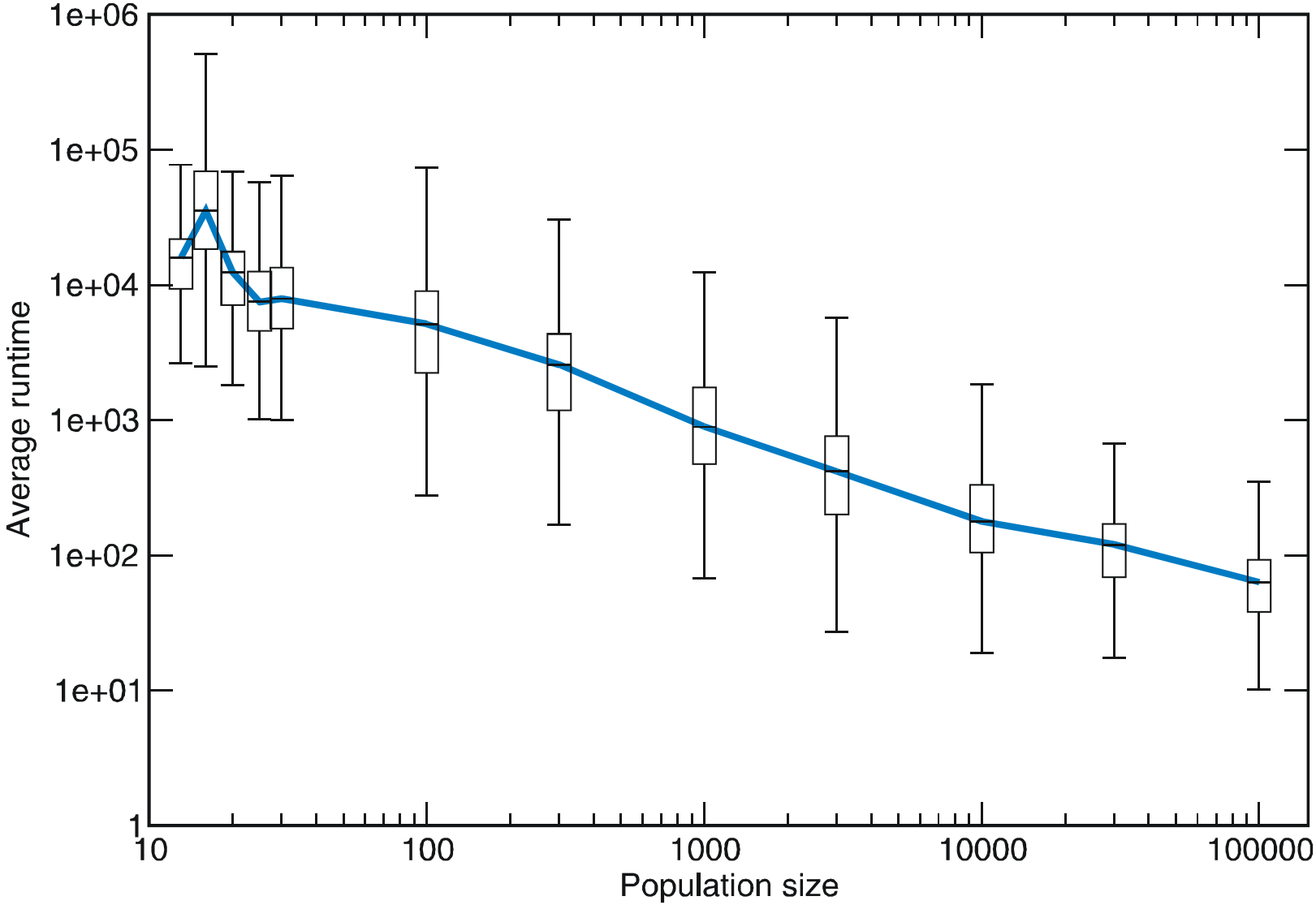
The neutral network of the tRNA clover leaf in **GC** sequence space is not connected, whereas to the corresponding neutral network in **AUGC** sequence space is close to the connectivity threshold, λ_{cr}. Here, both inverse folding and optimization in the flow reactor are much more effective than with **GC** sequences.

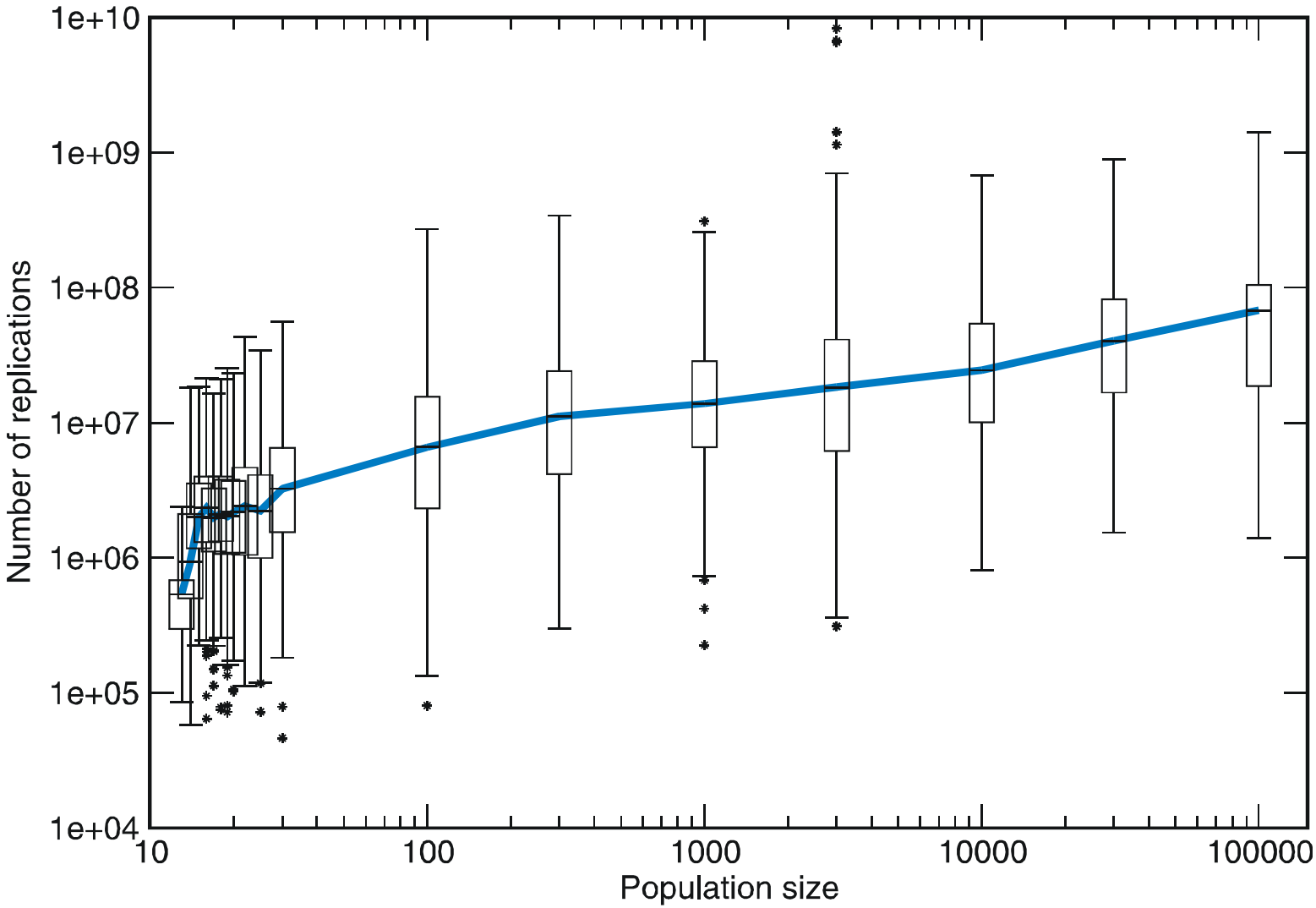
The hardness of the structure optimization problem depends on the connectivity of neutral networks.


Target
Initial state
Replication, mutation and dilution
Extinction



Probability of a single trajectory to reach the target structure


1e+06
1e+05
1e+04
1e+03
1e+02
1e+01
1
Average runtime
10
100
1000
10000
100000
Population size


1e+10
1e+09
1e+08
1e+07
1e+06
1e+05
1e+04
Number of replications
10
100
1000
10000
100000
Population size

A ribozyme switch

E.A.Schultes, D.B.Bartel, Science **289** (2000), 448-452

minus the background levels observed in the HSP in the control (Sar1-GDP–containing) incubation that prevents COPII vesicle formation. In the microsome control, the level of p115-SNARE associations was less than 0.1%.
46. C. M. Carr, E. Grote, M. Munson, F. M. Hughson, P. J. Novick, *J. Cell Biol.* **146**, 333 (1999).
47. C. Ungermann, B. J. Nichols, H. R. Pelham, W. Wickner, *J. Cell Biol.* **140**, 61 (1998).
48. E. Grote and P. J. Novick, *Mol. Biol. Cell* **10**, 4149 (1999).
49. P. Uetz *et al.*, *Nature* **403**, 623 (2000).
50. GST-SNARE proteins were expressed in bacteria and purified on glutathione-Sepharose beads using standard methods. Immobilized GST-SNARE protein (0.5 μM) was incubated with rat liver cytosol (20 mg) or purified recombinant p115 (0.5 μM) in 1 ml of NS buffer containing 1% BSA for 2 hours at 4°C with rotation. Beads were briefly spun (3000 rpm for 10 s) and sequentially washed three times with NS buffer and three times with NS buffer supplemented with 150 mM NaCl. Bound proteins were eluted three times in 50 μl of 50 mM tris-HCl (pH 8.5), 50 mM reduced glutathione, 150 mM NaCl, and 0.1% Triton X-100 for 15 min at 4°C with intermittent mixing, and elutes were pooled. Proteins were precipitated by MeOH/CH_3Cl and separated by SDS–polyacrylamide gel electrophoresis (PAGE) followed by immunoblotting using p115 mAb 13F12.
51. V. Rybin *et al.*, *Nature* **383**, 266 (1996).
52. K. G. Hardwick and H. R. Pelham, *J. Cell Biol.* **119**, 513 (1992).
53. A. P. Newman, M. E. Groesch, S. Ferro-Novick, *EMBO J.* **11**, 3609 (1992).
54. A. Spang and R. Schekman, *J. Cell Biol.* **143**, 589 (1998).
55. M. F. Rexach, M. Latterich, R. W. Schekman, *J. Cell Biol.* **126**, 1133 (1994).
56. A. Mayer and W. Wickner, *J. Cell Biol.* **136**, 307 (1997).
57. M. D. Turner, H. Plutner, W. E. Balch, *J. Biol. Chem.* **272**, 13479 (1997).
58. A. Price, D. Seals, W. Wickner, C. Ungermann, *J. Cell Biol.* **148**, 1231 (2000).
59. X. Cao and C. Barlowe, *J. Cell Biol.* **149**, 55 (2000).
60. G. G. Tall, H. Hama, D. B. DeWald, B. F. Horazdovsky, *Mol. Biol. Cell* **10**, 1873 (1999).
61. C. G. Burd, M. Peterson, C. R. Cowles, S. D. Emr, *Mol. Biol. Cell* **8**, 1089 (1997).
62. M. R. Peterson, C. G. Burd, S. D. Emr, *Curr. Biol.* **9**, 159 (1999).
63. M. G. Waters, D. O. Clary, J. E. Rothman, *J. Cell Biol.* **118**, 1015 (1992).
64. D. M. Walter, K. S. Paul, M. G. Waters, *J. Biol. Chem.* **273**, 29565 (1998).
65. N. Hui *et al.*, *Mol. Biol. Cell* **8**, 1777 (1997).
66. T. E. Kreis, *EMBO J.* **5**, 931 (1986).
67. H. Plutner, H. W. Davidson, J. Saraste, W. E. Balch, *J. Cell Biol.* **119**, 1097 (1992).
68. D. S. Nelson *et al.*, *J. Cell Biol.* **143**, 319 (1998).
69. We thank G. Waters for p115 cDNA and p115 mAbs; G. Warren for p97 and p47 antibodies; R. Scheller for rbet1, membrin, and sec22 cDNAs; H. Plutner for excellent technical assistance; and P. Tan for help during the initial phase of this work. Supported by NIH grants GM 33301 and GM42336 and National Cancer Institute grant CA58689 (W.E.B.), a NIH National Research Service Award (B.D.M.), and a Wellcome Trust International Traveling Fellowship (B.B.A.).

20 March 2000; accepted 22 May 2000

One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds

Erik A. Schultes and David P. Bartel*

We describe a single RNA sequence that can assume either of two ribozyme folds and catalyze the two respective reactions. The two ribozyme folds share no evolutionary history and are completely different, with no base pairs (and probably no hydrogen bonds) in common. Minor variants of this sequence are highly active for one or the other reaction, and can be accessed from prototype ribozymes through a series of neutral mutations. Thus, in the course of evolution, new RNA folds could arise from preexisting folds, without the need to carry inactive intermediate sequences. This raises the possibility that biological RNAs having no structural or functional similarity might share a common ancestry. Furthermore, functional and structural divergence might, in some cases, precede rather than follow gene duplication.

Related protein or RNA sequences with the same folded conformation can often perform very different biochemical functions, indicating that new biochemical functions can arise from preexisting folds. But what evolutionary mechanisms give rise to sequences with new macromolecular folds? When considering the origin of new folds, it is useful to picture, among all sequence possibilities, the distribution of sequences with a particular fold and function. This distribution can range very far in sequence space (*1*). For example, only seven nucleotides are strictly conserved among the group I self-splicing introns, yet secondary (and presumably tertiary) structure within the core of the ribozyme is preserved (*2*). Because these disparate isolates have the same fold and function, it is thought that they descended from a common ancestor through a series of mutational variants that were each functional. Hence, sequence heterogeneity among divergent isolates implies the existence of paths through sequence space that have allowed neutral drift from the ancestral sequence to each isolate. The set of all possible neutral paths composes a "neutral network," connecting in sequence space those widely dispersed sequences sharing a particular fold and activity, such that any sequence on the network can potentially access very distant sequences by neutral mutations (*3–5*).

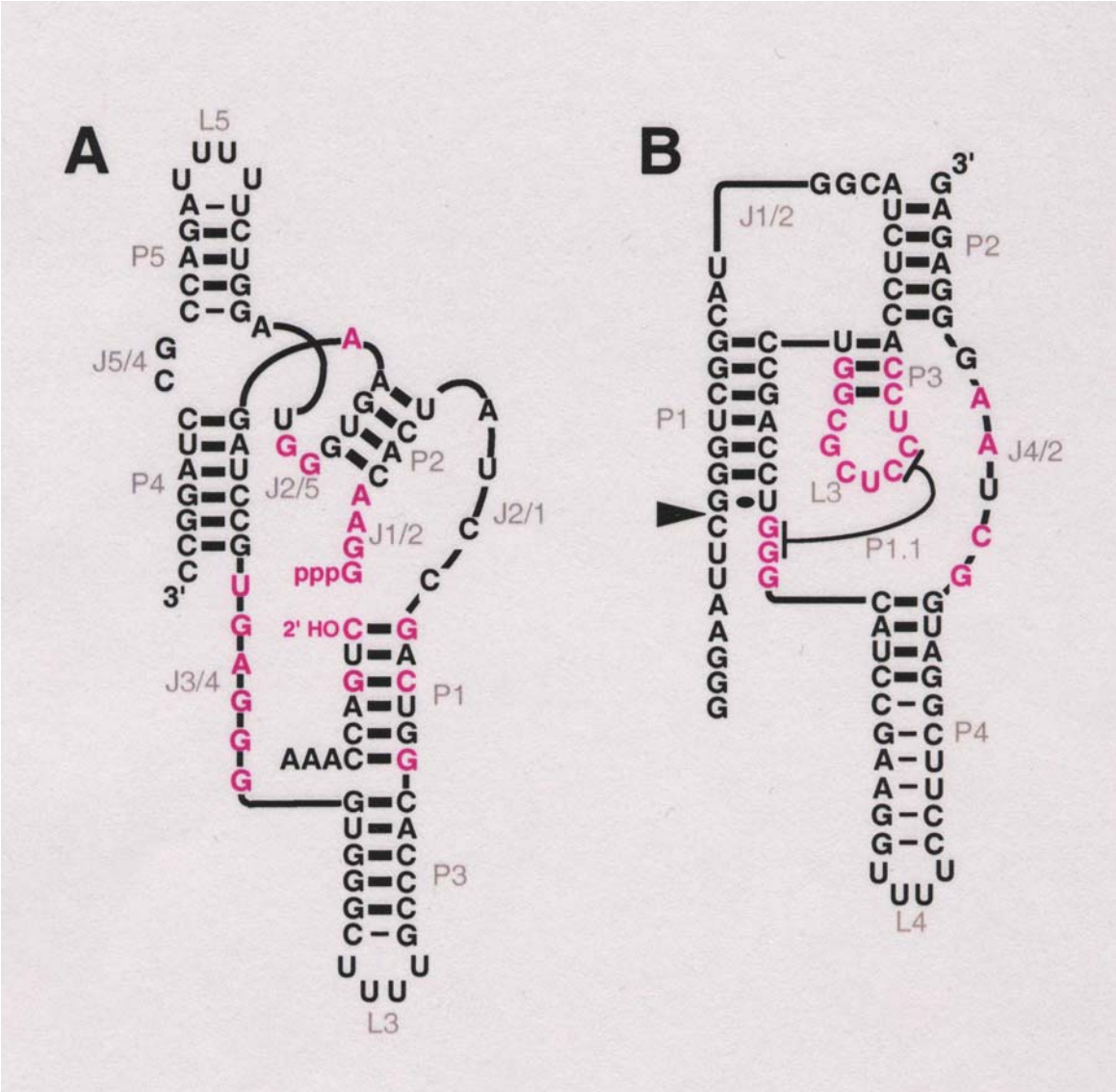
Theoretical analyses using algorithms for predicting RNA secondary structure have suggested that different neutral networks are interwoven and can approach each other very closely (*3*, *5–8*). Of particular interest is whether ribozyme neutral networks approach each other so closely that they intersect. If so, a single sequence would be capable of folding into two different conformations, would have two different catalytic activities, and could access by neutral drift every sequence on both networks. With intersecting networks, RNAs with novel structures and activities could arise from previously existing ribozymes, without the need to carry nonfunctional sequences as evolutionary intermediates. Here, we explore the proximity of neutral networks experimentally, at the level of RNA function. We describe a close apposition of the neutral networks for the hepatitis delta virus (HDV) self-cleaving ribozyme and the class III self-ligating ribozyme.

In choosing the two ribozymes for this investigation, an important criterion was that they share no evolutionary history that might confound the evolutionary interpretations of our results. Choosing at least one artificial ribozyme ensured independent evolutionary histories. The class III ligase is a synthetic ribozyme isolated previously from a pool of random RNA sequences (*9*). It joins an oligonucleotide substrate to its 5′ terminus. The prototype ligase sequence (Fig. 1A) is a shortened version of the most active class III variant isolated after 10 cycles of in vitro selection and evolution. This minimal construct retains the activity of the full-length isolate (*10*). The HDV ribozyme carries out the site-specific self-cleavage reactions needed during the life cycle of HDV, a satellite virus of hepatitis B with a circular, single-stranded RNA genome (*11*). The prototype HDV construct for our study (Fig. 1B) is a shortened version of the antigenomic HDV ribozyme (*12*), which undergoes self-cleavage at a rate similar to that reported for other antigenomic constructs (*13*, *14*).

The prototype class III and HDV ribozymes have no more than the 25% sequence identity expected by chance and no fortuitous structural similarities that might favor an intersection of their two neutral networks. Nevertheless, sequences can be designed that simultaneously satisfy the base-pairing requirements

Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

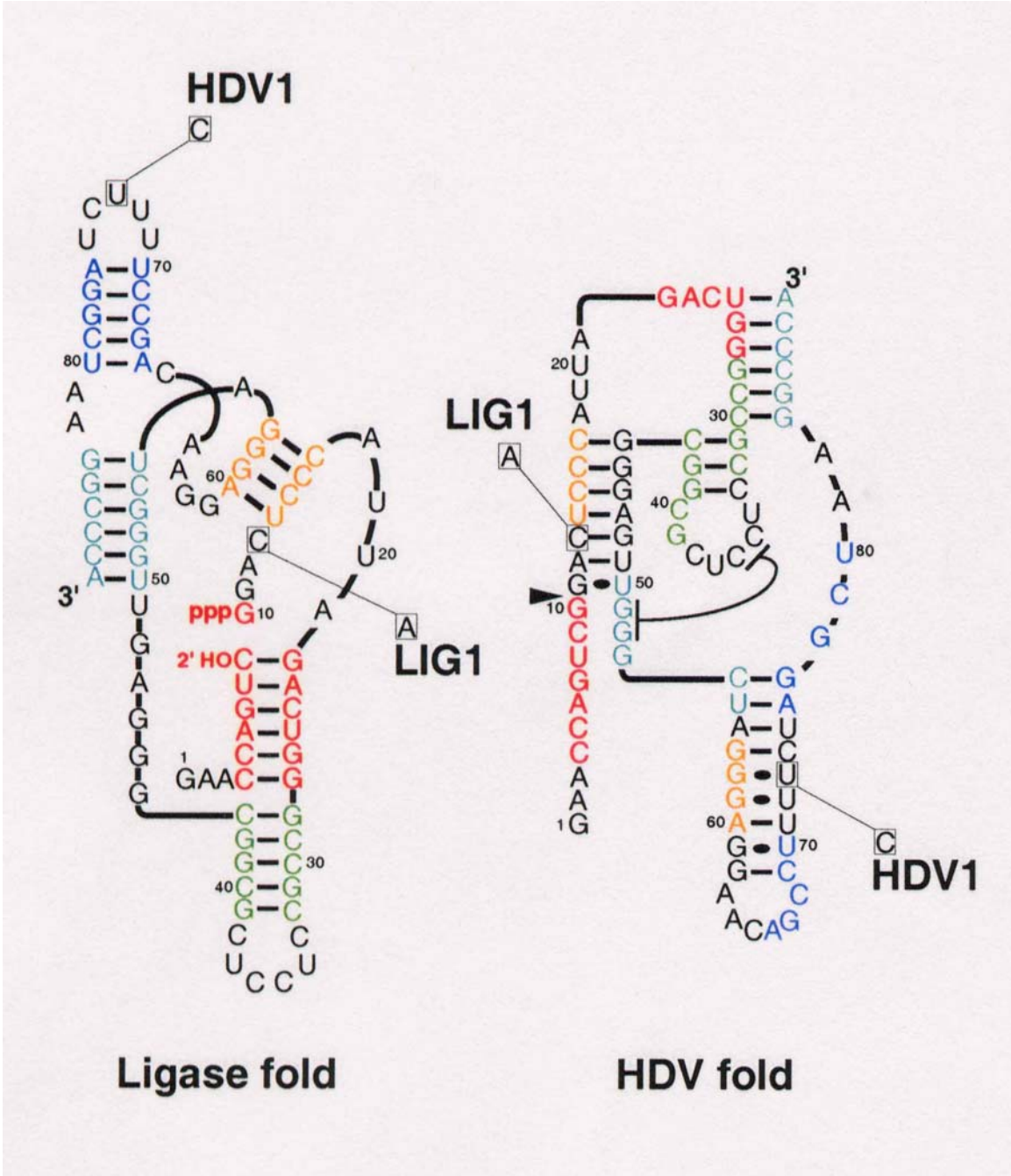
*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

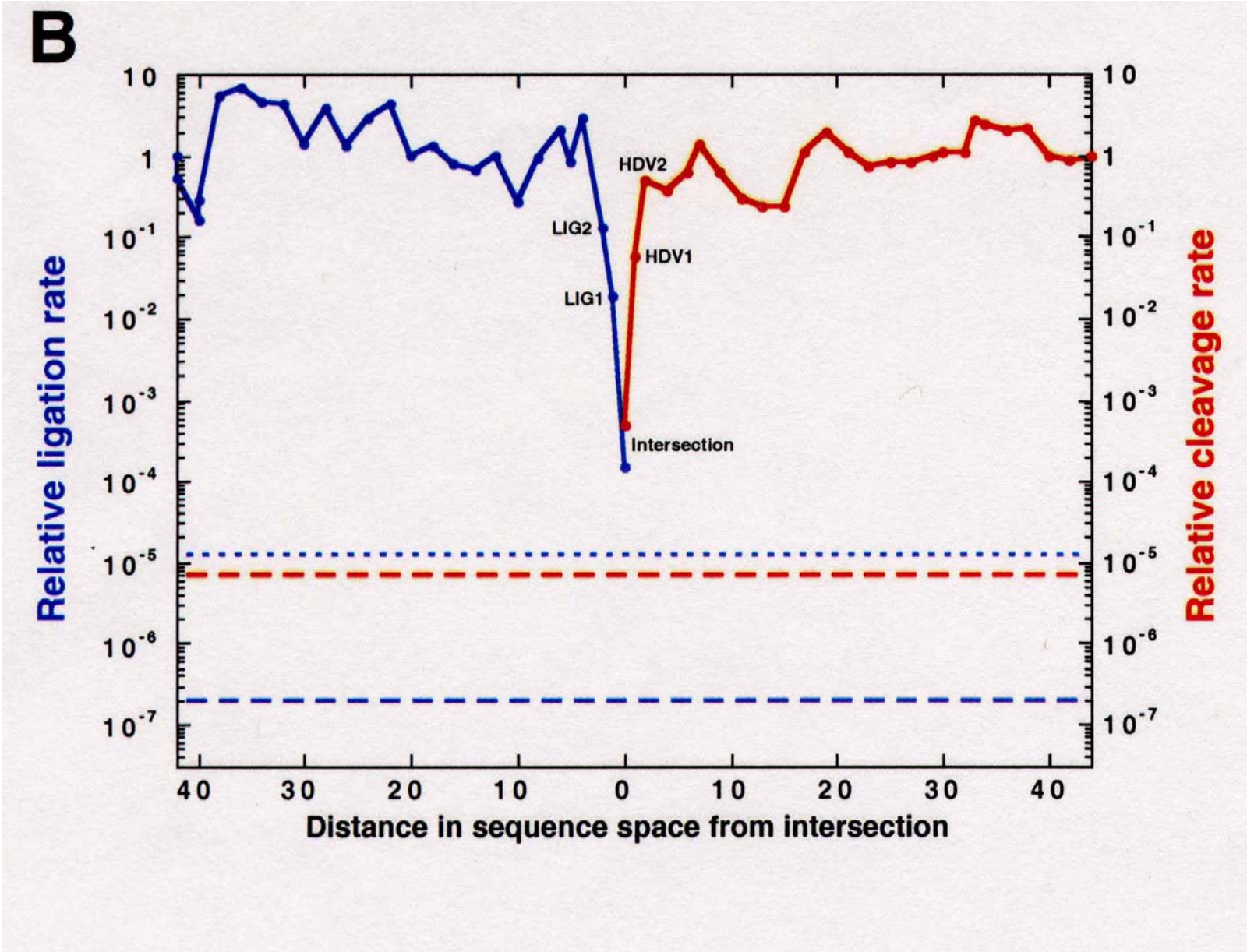


Two ribozymes of chain lengths n = 88 nucleotides: An artificial ligase (**A**) and a natural cleavage ribozyme of hepatitis-δ-virus (**B**)

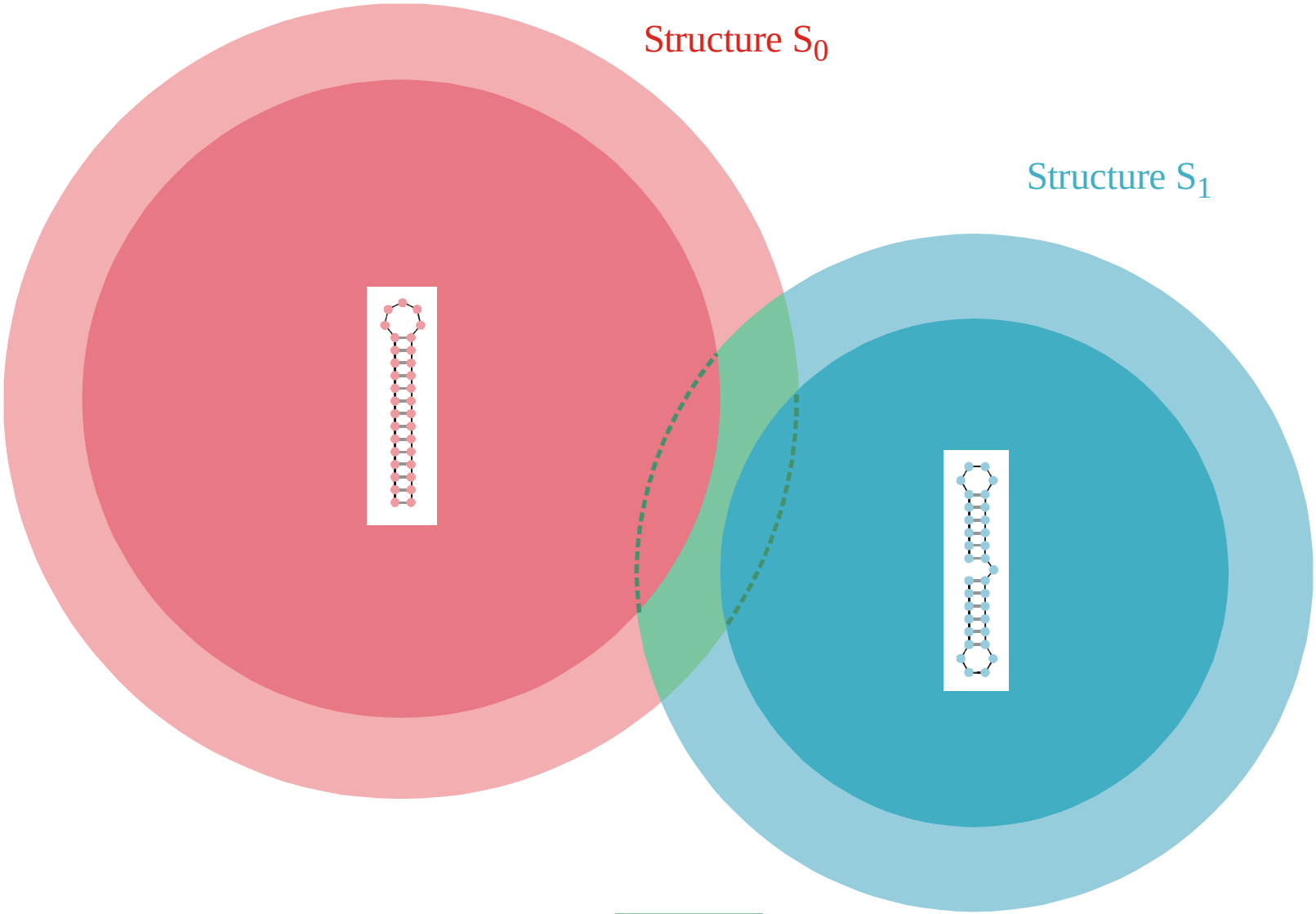
The sequence at the ***intersection***:

An RNA molecules which is 88 nucleotides long and can form both structures





Two neutral walks through sequence space with conservation of structure and catalytic activity

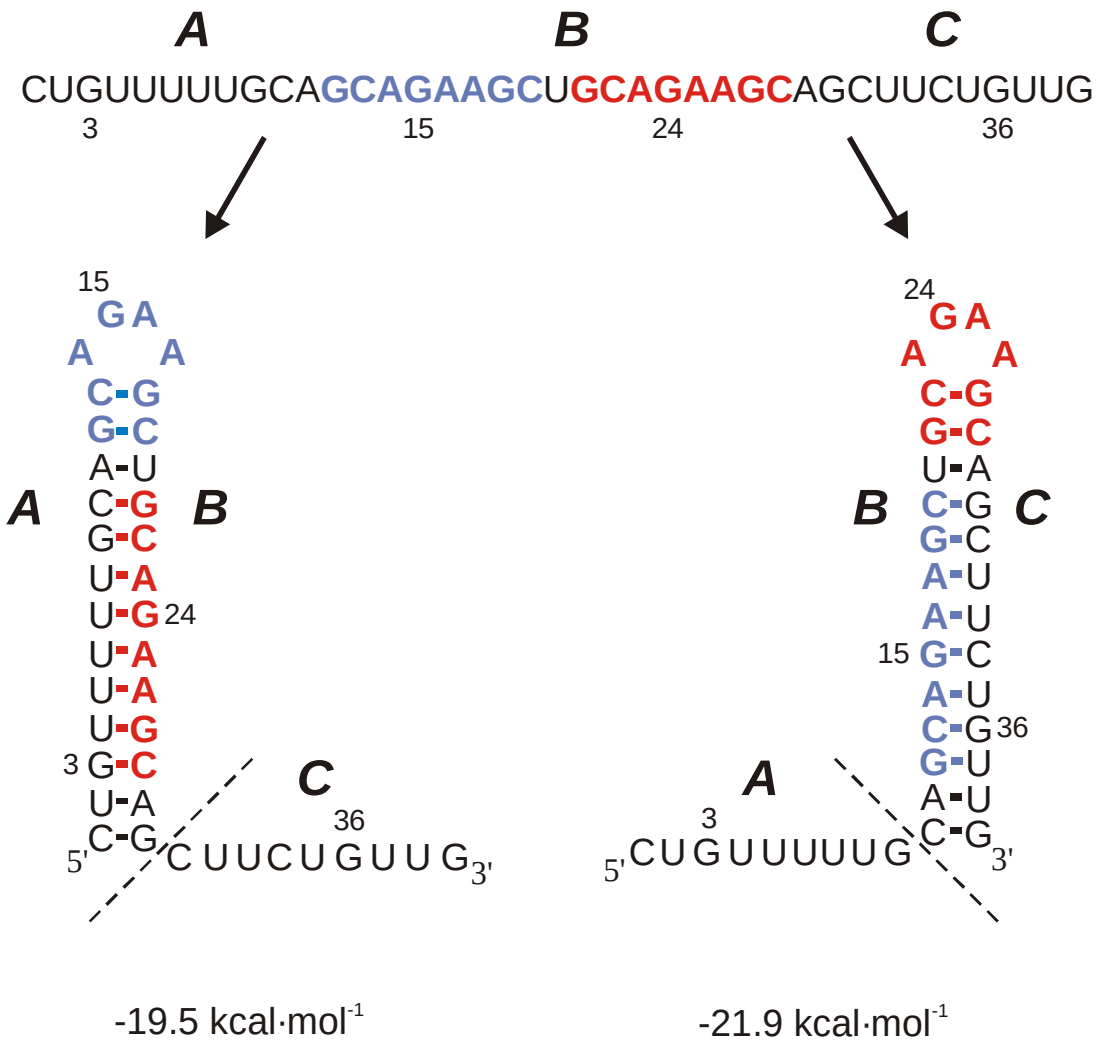


Intersection of two compatible sets: $C_0 \cap C_1$

The intersection of two compatible sets is always non empty: $C_0 \cap C_1 \notin \varnothing$

J. H. A. Nagel, C. Flamm, I. L. Hofacker, K. Franke, M. H. de Smit, P. Schuster, and C. W. A. Pleij. ***Structural parameters affecting the kinetic competition of RNA hairpin formation***, in press 2004.

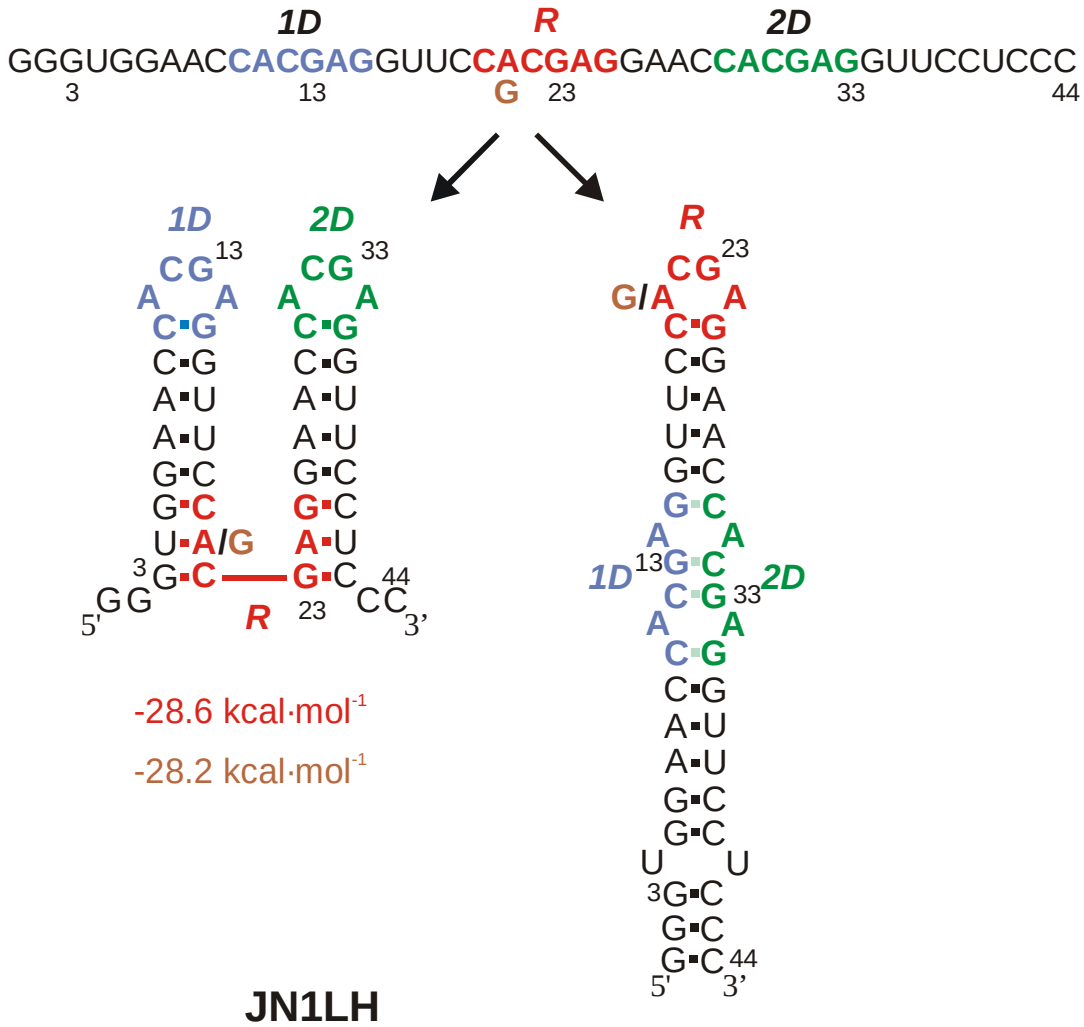
J. H. A. Nagel, J. Møller-Jensen, C. Flamm, K. J. Öistämö, J. Besnard, I. L. Hofacker, A. P. Gultyaev, M. H. de Smit, P. Schuster, K. Gerdes and C. W. A. Pleij. ***The refolding mechanism of the metastable structure in the 5'-end of the*** hok ***mRNA of plasmid*** R1, submitted 2004.



JN2C

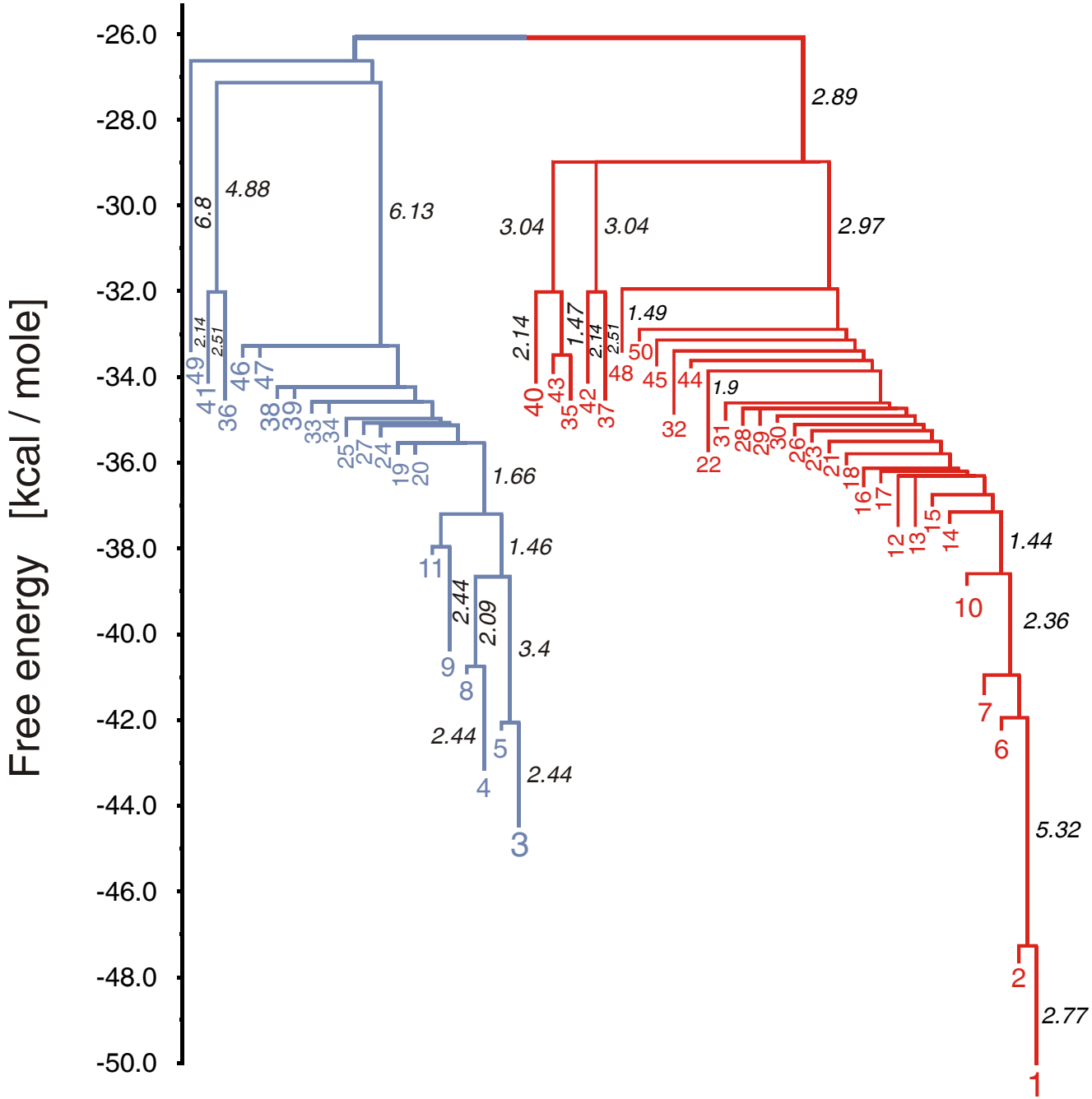
J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke,
M.H. de Smit, P. Schuster, and C.W.A. Pleij.

Structural parameters affecting the kinetic competition of RNA hairpin formation, in press 2004.



J.H.A. Nagel, C. Flamm, I.L. Hofacker, K. Franke, M.H. de Smit, P. Schuster, and C.W.A. Pleij.

Structural parameters affecting the kinetic competition of RNA hairpin formation, in press 2004.



J1LH barrier tree

Riboswitches

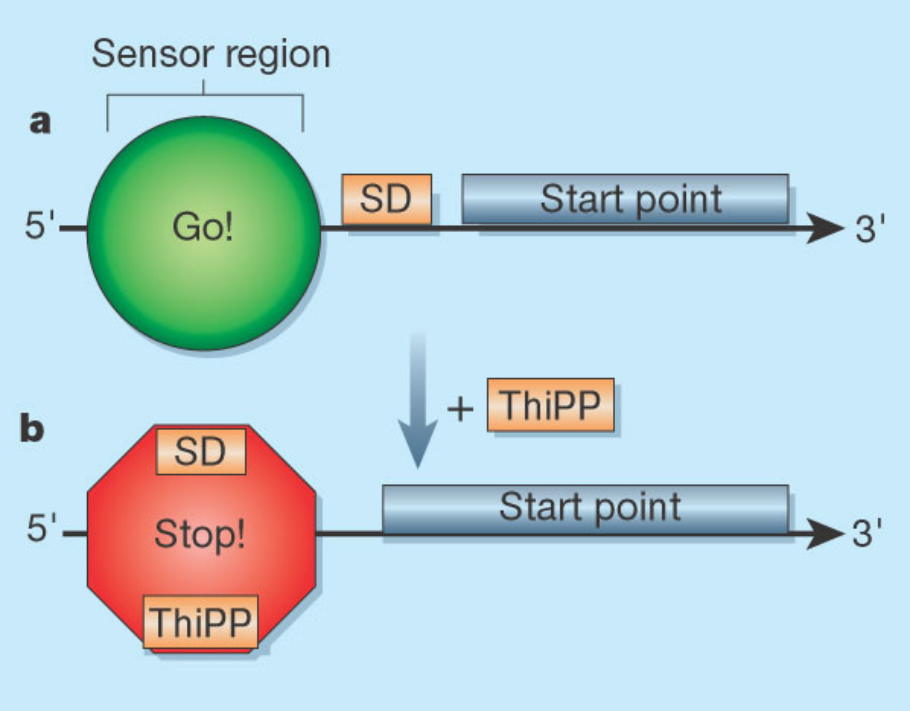
Jord H. A. Nagel and Cornelius W. A. Pleij. ***Self-induced structural switches in RNA***. Biochimie **84** (2002), 913-923

Wade Winkler, Ali Nahvi, and Ronald R. Breaker. ***Thiamine derivatives bind messenger RNA directly to regulate bacterial gene expression***. Nature **419** (2002), 952-956

Ronald Micura and Claudia Höbartner. ***On Secondary Structure Rearrangements and Equilibria of Small RNAs***. Nature **419** (2002), 952-956

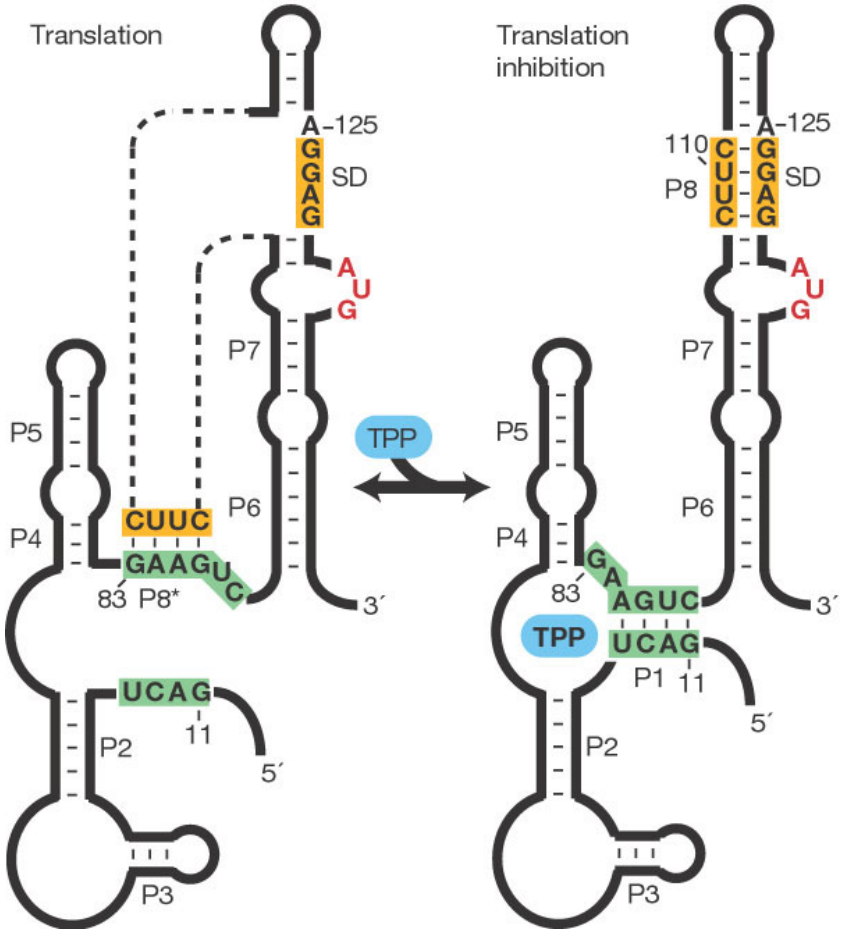
Alexey G. Vitreschak, Dimitry A. Rodionov, Andrey A. Mironov, and Mikhail S. Gefland. ***Riboswitches: The oldest mechanism for the regulation of gene expression?*** Trends in Genetics **20** (2004), 44-50

Jeffrey E. Barrick, Keith A. Corbino, Wade C. Winkler, Ali Nahvi, Maumita Mandal, Jennifer Collins, Mark Lee, Adam Roth, Narasimhan Sundarasan, Inbal Jona, J. Kenneth Wickiser, and Ronald R. Breaker. ***New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control***. Proc.Natl.Acad.Sci.USA **101** (2004), 6421-6426



RNA as metabolite sensor and translation regulator: The figure represents part of a messenger RNA that encodes a bacterial protein involved in the biosynthesis of vitamin B_1 (thiamine pyrophosphate, ThiPP). This part contains a region involved in sensing ThiPP; the so-called Shine–Dalgarno (SD) sequence, which is recognized by the ribosome; and a sequence that marks where the enzyme-encoding portion of the mRNA begins.
a, Winkler *et al.*[1] have found that in the absence of ThiPP, the sensor region adopts a conformation that exposes the Shine–Dalgarno sequence. This would
allow the ribosome to bind and begin translation.
b, Binding to ThiPP causes the sensor region to change shape, obscuring the Shine–Dalgarno sequence. Thus ThiPP controls the production of one of its biosynthetic enzymes directly via a sequence within the enzyme-encoding mRNA.

Jack W. Szostak, RNA gets a grip on translation. *Nature* **419**:890-891, 2002



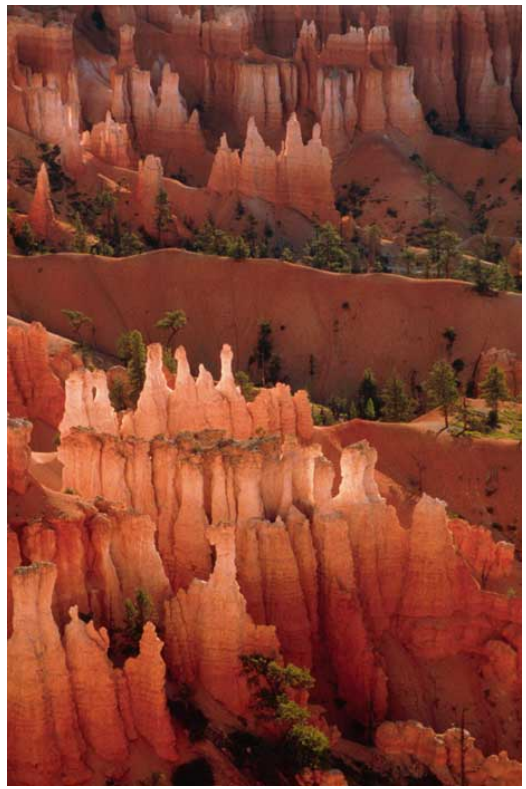
Schematic representation of the proposed mechanism for TPP-dependent deactivation of *thiM* translation: In the absence of TPP, the P8* pairing is formed between the anti-SD element and the anti-anti-SD element. This conformation permits the SD sequence to interact with the ribosome, and thus translation proceeds. In the presence of TPP (blue), the obligate formation of the P1 stem sequesters a portion of the anti-anti-SD element, and therefore the complete P8 stem also forms. This precludes ribosome access to the SD element, which inhibits translation. Complementary sequence elements that form P1 and P8 are depicted in green and orange, respectively.

Wade Winkler, Ali Nahvi, and Ronald R. Breaker, Thiamine derivatives bind messenger RNA directly to regulate bacterial gene expression. *Nature* **419**:952-956, 2002

Mount Fuji

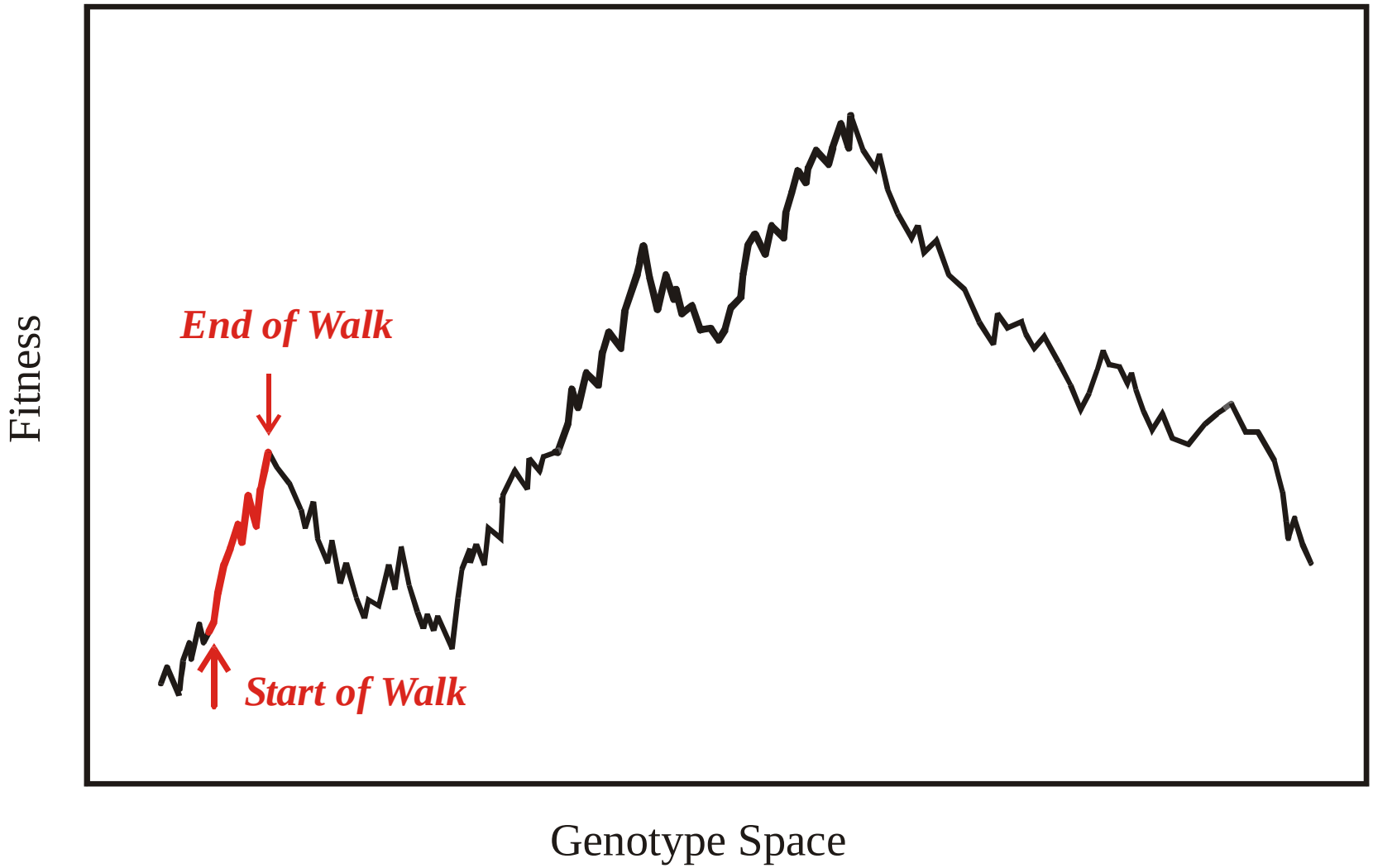
Example of a smooth landscape on Earth

Dolomites

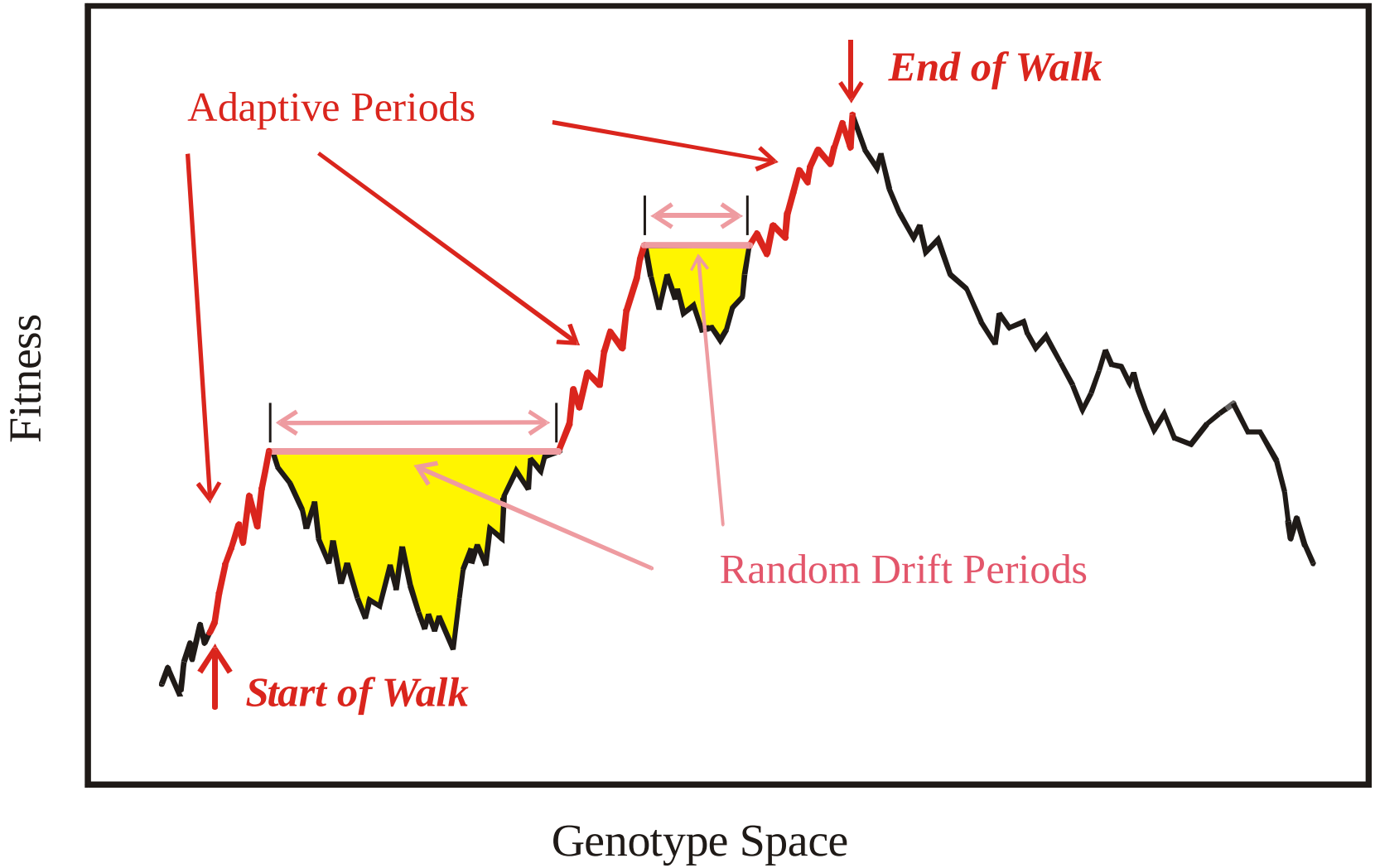


Bryce Canyon

Examples of rugged landscapes on Earth

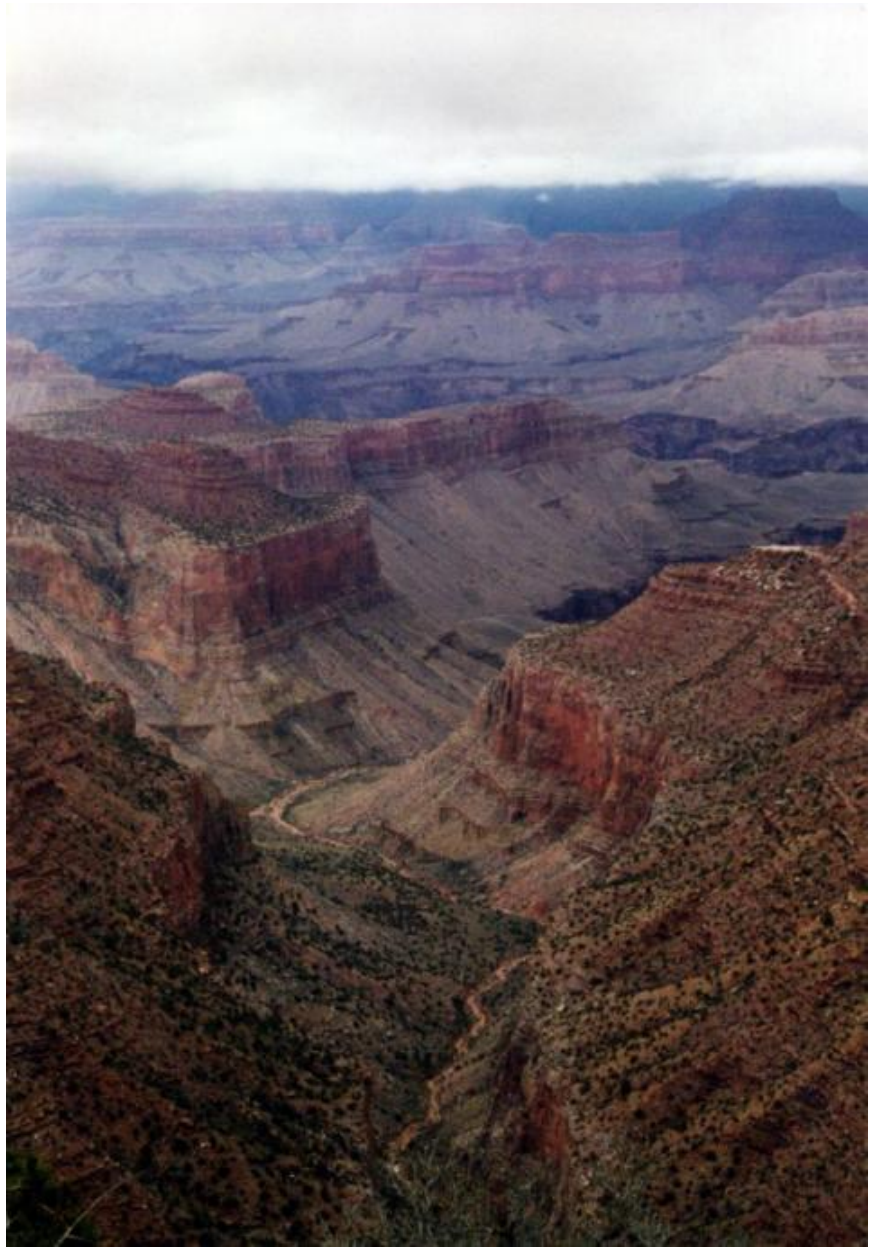


Evolutionary optimization in absence of neutral paths in sequence space



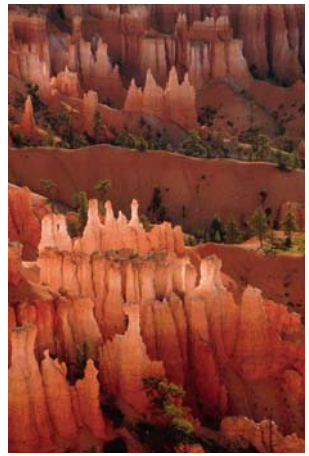
Evolutionary optimization including neutral paths in sequence space

Grand Canyon

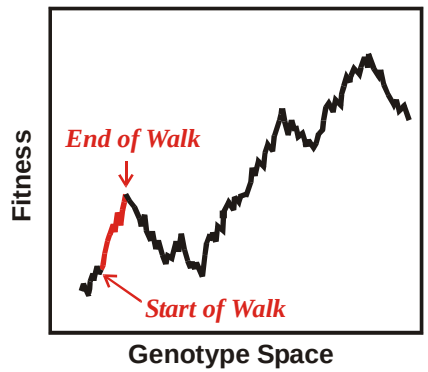


Example of a landscape on Earth with ‘neutral’ ridges and plateaus

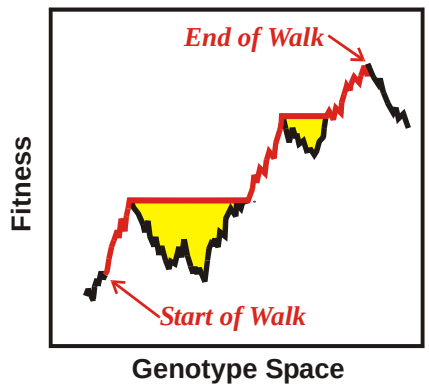
- Conformational and mutational landscapes of biomolecules as well as fitness landscapes of evolutionary biology are **rugged**.



- **Adaptive** or **non-descending walks** on rugged landscapes end commonly at one of the low lying local maxima.



- Selective neutrality in the form of **neutral networks** plays an active role in evolutionary optimization and enables populations to reach high local maxima or even the global optimum.



Acknowledgement of support

Fonds zur Förderung der wissenschaftlichen Forschung (FWF)

Projects No. 09942, 10578, 11065, 13093, 13887, and 14898

Jubiläumsfonds der Österreichischen Nationalbank

Project No. Nat-7813

European Commission: Project No. EU-980189

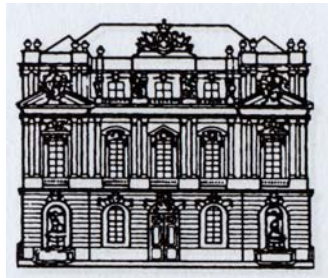
Siemens AG, Austria

Österreichische Akademie der Wissenschaften

Universität Wien

The software for producing RNA movies was developed by
Robert Giegerich and coworkers at the Universität Bielefeld

Universität Wien



Österreichische Akademie der Wissenschaften

Coworkers

Universität Wien

Walter Fontana, Harvard Medical School, Boston, MA

Christian Reidys, Christian Forst, Los Alamos National Laboratory, NM

Peter Stadler, **Bärbel Stadler,** Universität Leipzig, GE

Ivo L.Hofacker, Christoph Flamm, Andreas Svrček-Seiler, Universität Wien, AT

Kurt Grünberger, Stefan Müller, Stefanie Widder, Stefan Wuchty,
Andreas Wernitznig, **Michael Kospach, Jan Cupal, Stefan Bernhard,**
Lukas Endler, Ulrike Langhammer, Ulrike Mückstein, Universität Wien AT

Ulrike Göbel, Walter Grüner, Stefan Kopp, Jaqueline Weber
Institut für Molekulare Biotechnologie, Jena, GE

Web-Page for further information:

http://www.tbi.univie.ac.at/~pks